

VISIT...

LANZAROTE
Caliente.COM

Protein Structure: Geometry, Topology and Classification

William R. Taylor, Alex C. W. May,
Nigel P. Brown[†] and András Aszódi[‡]

Division of Mathematical Biology,
National Institute for Medical Research,
The Ridgeway, Mill Hill, London NW7 1AA, U.K.

[†] currently at: Protein Design Group, Centro Nacional de Biotecnología,
Campus Universidad Autónoma, Cantoblanco, 28049 Madrid, Spain.

[‡] currently at: Novartis Forschungsinstitut GmbH, Brunnerstrasse 59,
A-1235 Vienna, Austria.

March 8, 2001

Contents

I	Introduction	6
1	Prologue	7
1.1	Scope and Aims	7
1.2	Why Proteins?	8
1.2.1	Catching a Demon	8
1.2.2	Origins	8
1.3	Outline of the Work	11
2	Basic Principles of Protein Structure	12
2.1	The shapes and sizes of proteins	12
2.1.1	Fibrous proteins	13
2.1.2	Globular proteins	13
2.1.3	Membrane proteins	13
2.2	The hydrophobic core	14
2.3	Secondary structure	14
2.4	Packed layers	16
2.4.1	All- α proteins	16
2.4.2	All- β proteins	16
2.4.3	α - β proteins	17
2.5	Barrel structures and β -helices	17
2.6	Protein Topology	19
2.7	Domain structure	19
II	Protein Structure Comparison and Classification	23
3	Overview of Comparison Methods	24
3.1	Challenges for Structure Comparison Methods	24
3.2	Degrees of Difficulty	25
3.3	Different Approaches	26
3.3.1	Comparison Power	26
3.3.2	Feature or Relationship	26
3.3.3	Hybrid methods	27
3.4	Dynamic Programming	27
3.4.1	The basic evolutionary model	27
3.4.2	Sequence Alignment	29
3.4.3	Gap-penalty	29
3.4.4	Structure Biased Gap-penalties	31

4	Early and Simple Approaches	31
4.0.5	Manual and semi-automatic methods	32
4.0.6	Fragment based methods	32
4.1	Comparing Feature Strings	33
4.1.1	Residue level	33
4.1.2	Backbone-fragment level	34
4.1.3	Secondary structure level	34
5	3D Methods without dynamic programming	34
5.1	Distance-matrix matching	35
5.1.1	Early attempts	35
5.1.2	The DALI method	36
5.1.3	Backbone fragment methods	36
5.2	Secondary structure graph-matching	37
5.3	Geometric-hashing approach	38
6	3D Methods using Dynamic Programming	40
6.1	Using structural superposition	40
6.2	Using the relationships of internal features	41
6.2.1	The COMPARE program	41
6.2.2	The SSAP program	42
6.3	Iterated Double Dynamic Programming	44
6.3.1	Double Dynamic Programming	44
6.3.2	Selection and Iteration	45
6.3.3	Sampling alternate alignments	45
7	Assessment of Significance	47
7.1	Score distributions from known structures	47
7.2	Random structural models	48
7.3	Randomised alignment models	48
7.4	Scoring and biological significance	49
7.5	Examples	50
7.5.1	Distant globin similarities	50
7.5.2	Assessment against chain reversal model	52
8	Protein Structure Classification	54
8.1	Introduction	54
8.1.1	Practical applications	54
8.1.2	Genome applications	55
8.2	Practical approaches to classification	55
8.2.1	Automated approaches to classification	56
8.3	Organisation of the classifications	57
8.3.1	The unit of classification	57

8.3.2	Hierarchical organisation	57
8.3.3	Hierarchical classification	58
8.4	Remaining Problems	58
8.4.1	What questions does classification help us to answer? . . .	58
8.4.2	Questions raised by classification	59
8.4.3	Future prospects	60
III	Geometric Abstractions and Topology	61
9	Simplified Geometries	62
9.1	Structure Representations	62
9.1.1	From bonds to cartoons	62
9.1.2	From 3-D to 2-D	62
10	Stick Representation	64
10.1	Secondary structure line-segments	64
10.1.1	Problems with current criteria	64
10.1.2	Line segments from inertial axes	65
10.1.3	Dynamic programming solution	66
10.1.4	‘Continuous’ secondary structure types	67
11	Ideal Forms	67
11.1	Layer Architectures	70
11.1.1	$\alpha/\beta/\alpha$ layers	70
11.1.2	β/β layers	71
11.1.3	β/α -barrel proteins	71
11.1.4	All- α proteins	71
11.1.5	Transmembrane models	71
11.2	Stick-figure comparisons	73
11.2.1	Angle and Distance matching	73
11.2.2	Finding the best match	73
11.2.3	Evaluation using SAP	76
11.2.4	Nested solutions	76
11.3	Classification using ideal stick forms	78
11.3.1	A periodic table of proteins	78
12	Fold Combinatorics	80
12.0.2	Motif incorporation	80
12.1	Evaluating folds	81

13 Protein Topology	84
13.1 Introduction	84
13.2 Chemical topology	84
13.3 Polymer topology	85
13.3.1 Bond direction	86
13.3.2 Linear polymers	86
13.3.3 Branching polymers	86
13.3.4 Circular polymers	86
13.4 True Topology of Proteins	87
13.4.1 Disulfide bridges	87
13.4.2 Other cross-links	89
13.5 Pseudo-Topology of Proteins	89
13.5.1 Topology of weak links in proteins	89
13.5.2 Topology of ‘circular’ proteins	90
13.5.3 ‘Topology’ of open chains	91
14 Symmetry	97
14.1 Structural origins of fold symmetries	97
14.1.1 $\beta\alpha$ -class	97
14.1.2 $\beta\beta$ -class	97
14.1.3 $\alpha\alpha$ -class	98
14.2 Evolutionary origins of fold symmetries	98
14.3 Conclusions	100

Part I

Introduction

The ultimate rationale behind all purposeful structures and behaviour of living beings is embodied in the sequence of residues of nascent polypeptide chains — the precursors of the folded proteins which in biology play the role of Maxwell's demons. In a very real sense it is at this level of organisation that the secret of life (if there is one) is to be found. If we could not only determine these sequences but also pronounce the law by which they fold, then the secret of life would be found — the ultimate rationale discovered!

Jaques Monod (1970)

from *Chance and Necessity*

loosely translated from the French (and Latin).

1 Prologue

1.1 Scope and Aims

Proteins are the main essential active agents in biochemistry: without them almost none of the metabolic processes that we associate with life would take place. Consequently, most reviews of proteins concentrate on these catalytic abilities: on their chemical kinetics, interactions and the detailed stereochemical arrangement of the catalytic groups that allow catalysis (or binding) of substrate and other macromolecules. From this biochemical viewpoint, the overall structure of the protein (which is much larger than the active-site) is viewed as a relatively uninteresting supporting scaffold for the chemistry. In this review, however, proteins will be viewed from a different angle — indeed, their biology and chemistry will be completely ignored. Instead, their overall structure will form the central topic and within this, an emphasis will be placed on abstracting an overview rather than concentrating on chemical or structural details. The underlying theme of the work is: “why do proteins adopt the forms that we see?” leading to the supplementary question: “do the proteins we know represent a fraction or a full sample of the possible forms?”. The answers to these questions are not only of interest from a structural/biochemical viewpoint but also have implications for our ideas of molecular evolution and the origin of life.

The text of this work will be aimed at readers from the physical and mathematical sciences and, as such, will not rely on any significant biochemical knowledge on the part of the reader. Each topic will be fully explained from first principles with an emphasis on basic concepts rather than applications or occurrences. Much of the text will also concentrate on computational methods, again focusing on the basic algorithms rather than their application or implementation. As such, while essentially a review, little attempt has been made to provide an exhaustive coverage of the specialised literature. Rather, effort has been directed towards communicating ideas and methods that might have some resonance for those with a more physical background.

Many of the aspects of proteins that will be explored have been investigated by molecular biologists (such as ourselves) who have been enticed into more abstract areas. Along the way we have usually taken a pragmatic approach to each investigation, sometimes inventing new methods (which often turn out to be re-inventions) or ‘borrowing’ methods and approaches from other fields (especially physics). It is our hope in writing the current work, that some more specialised readers, perhaps having seen a frightening misapplication of their favourite method, might take-up the challenge and “do it properly”. There are also some problems discussed for which we, at least, see no way forward (or more generally, no satisfactory way forward). We hope that these topics might inspire consideration from a fresh (ideally, orthogonal) viewpoint and allow some new directions to be identified.

1.2 Why Proteins?

1.2.1 Catching a Demon

There are many large biological molecules, including: nucleic acids, carbohydrates, lipids and proteins. While each play a vital (and interesting) part in life, there is something special about proteins. From a physicist's point-of-view, the essence of this uniqueness might be captured by saying that, mechanically (if not thermodynamically), proteins are about as close as we can come to capturing a real-life Maxwell's Demon (Figure 1).

Of the components that make up life, almost all but proteins are relatively inert and are, generally, the substrates that are chopped and changed by the action of proteins. In doing this, proteins do not act using some abstract bulk property (as do lipids and carbohydrates) but are individual agents (rather like demons) that latch-onto their 'victims' (substrates) and cut and change them (sometimes even using the chemistry of sulphur). Indeed, when located across a lipid membrane, they are also quite good at opening and shutting trap-doors!

To a large extent, understanding the action of proteins is the key to understanding the spark of life itself and this has been stated quite explicitly in the quotation by Jaques Monod (one of the 'founding-fathers' of molecular biology) that opens this section. As indicated by Monod in the same quotation, proteins also occupy a unique position in the hierarchy of physical organisation: lying in a grey region between chemistry and biology. For a chemist, proteins are large complicated molecules that even polymer chemists would have difficulty in modelling. From the biological side, although any individual protein would not be considered to be alive, it does not take many of them (plus a bit of nucleic acid) before life-like behaviour begins to emerge. For example; some of the smallest viruses, such as HIV, which might be considered to be on the borderline of life, operates with only 10 different types of protein.

1.2.2 Origins

Before leaving these ideas about the nature of life it is interesting to consider some of the ideas concerning the origin of proteins. It is now generally accepted that, before the first living cells (just under four giga-years ago), 'life' — or rather the assemblies of self reproducing macromolecules — were ribonucleic acids (RNA). Circumstantial evidence for this can be found in 'relic' pieces of RNA that still hold a few of the most central functions in the processes of life: for example in the synthesis of proteins on the ribosome. In this 'RNA world' a single type of molecule performed both the functions of active (catalytic) agent and repository of its own description — the 'blueprint' from which further copies could be taken. The former function is a property of the folded molecule while the later is a property of the linear polymer sequence, and the two functions need not necessarily be compatible. One can imagine a situation in which, say, for

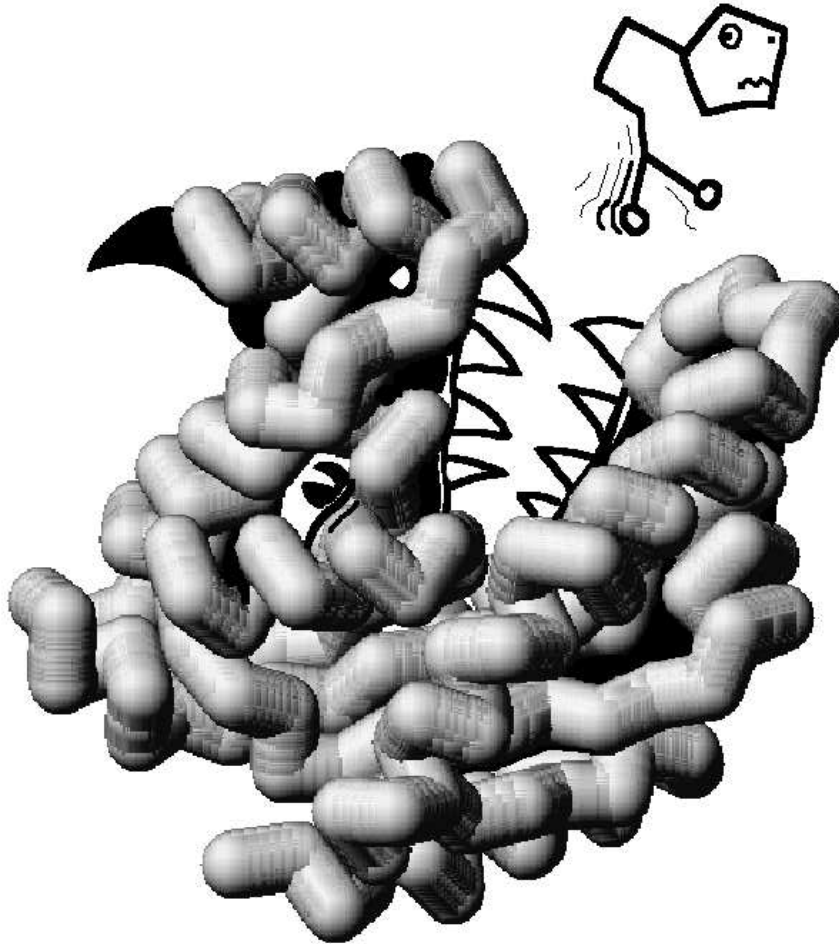


Figure 1: **A small enzyme approaches its substrate.** Against all thermodynamic reason, some people have likened proteins (such as this adenylate kinase molecule) to Maxwell's demons. The active (or catalytic) site of the molecule is indicated by elongated triangles.

more efficient catalysis, an extra chemical activity was needed at a particular point: however, a modification of the RNA structure to achieve this (such as the chemical modification of a part of the molecule) might leave it incapable of duplication or folding. It seems that RNA circumvented this problem by recruiting co-factors that could augment its chemical repertoire without compromising its ability to make copies of itself. Some of these cofactors were probably peptides and a development can be imagined in which the peptide cofactors became more complex as the functional rôle of the RNA diminished.

In this simple world, however, RNA would rely on the chance synthesis of suitable peptides which would limit both the size of the peptides and the number of these that could be involved with the RNA. This fundamental problem was overcome through the establishment of a synergistic loop in which the RNA was able to act as a template to guide the synthesis of the peptides that it needed. With the limitation on the chance synthesis of the right peptides now removed (or limited only by the fidelity in the translation of RNA into peptide sequences), the system was free to become much more complex¹. The details of how this key event in life became established are very vague but some plausible hypotheses are described in the opening chapters of “The RNA World” (Gesteland and Atkins, 1993). This transition marked the escape from the error-prone world of self-replicating macromolecules to a system with unlimited scope to control its own metabolism and replication. It also began the divergence of function: with peptides/proteins taking-over the active (catalytic) activity while RNA became more inert with its main function being now to encode proteins, which would then periodically help replicate the RNA itself.

From this state, the introduction of the third major molecular component of life — DNA — is almost incidental. With RNA free from most of its structural constraints and under strong evolutionary pressure to maintain the reproductive fidelity of the increasingly complex protein/RNA machine: in computer terms, a back-up facility was required. This was found in DNA, which is only a slightly modified form of RNA but has much greater stability — especially when ‘locked’ away in its famous double helical structure. This subsidiary rôle for DNA is maintained in all present day life: and although proteins can interact directly with DNA, there is no direct link from DNA to protein except via RNA intermediates.

It is interesting to note that this polarisation of function into active machine (protein) and inert blueprint (DNA) follows the logical requirements specified by von Neuman for a self replicating machine.

¹The short time-span between the impact that created the moon and the first cell have led some to suspect that there was not enough time for this complexity to develop on Earth.

1.3 Outline of the Work

Hopefully, the preceeding thoughts and speculations have proved to be sufficiently intriguing to persuade the less biologically-oriented reader that proteins are a fascinating topic and certainly one central to the understanding of life. In the following sections, we will leave these broader considerations and lay down some basic ground-work on protein structure so that all readers, irrespective of background, will have a common foundation on which some of the later more technical sections can build. As promised above, we will try to avoid the standard ‘biochemistry textbook’ approach to the topic.

From this base, the first major aspect to be considered will be the systematic comparison and classification of protein structure (Part II), progressing towards more abstract geometric representations of protein structure (Part III). These sections will include details and reviews of the methods that can be used for structure comparison and the degree to which these can be interpreted. For the physicist and mathematician, there should be interesting problems here involving the description of three-dimensional objects and the statistical significance of their comparison.

2 Basic Principles of Protein Structure

In this section the basic principles that determine protein structure will be reviewed. Although many aspects of these topics will be returned to in greater detail in the following sections below, it is better firstly to gain an overview of these together in one place rather than encounter important definitions scattered throughout the text. Further information on many of these topics, including greater biological background, can be found in Brändén and Tooze (1991) or Chothia (1984) for a review concentrating more on on packing.

2.1 The shapes and sizes of proteins

From a chemical viewpoint, proteins are linear hetropolymers. However, unlike most synthetic polymers, which are condensed from one or a few monomer units, proteins can draw on a mix of twenty different monomers. A further distinction is found in their organisation: while polymers are generally very large extended molecules forming a matrix (typically cross-linked as a gel), the majority of proteins fold as relatively small self-contained structures. These factors balance: although small (for a polymer), the variety of monomers gives an almost unlimited scope for the construction of different protein molecules. Perhaps the most remarkable feature of proteins, however, is the observation that each protein found in nature has a specific three-dimensional structure and that this structure is determined (effectively) only by the sequence of the monomers themselves. To give names to these parts: the monomer units are amino acids which condense with the formation of a peptide bond linking them: hence, the resulting chain is often referred to as a polypeptide. The linked amino acids are then referred to as residues: an odd name deriving from the stuff at the bottom of test-tubes when proteins were sequenced by chemical means in the early days of protein chemistry.

There is great variety in the structure of the twenty different (natural) amino acids but despite this, the variation (with one exception) is all confined to the side groups leaving a constant unit that polymerises into a regular backbone chain. (See Taylor (1986a) and Taylor (1999a) for some further discussion of amino acid properties). Furthermore, even though amino acids contain a chiral centre (on their α -carbon), only one enantiomer is used to make proteins. As we shall see below, this regularity in the polypeptide chain allows the formation of semi-regular substructures that are the building blocks of proteins. The polypeptide chain is also very flexible: although the peptide bond is not free to rotate, the two flanking bonds are, giving two reasonably free rotations for each residue.

2.1.1 Fibrous proteins

There is no (reasonable) physical limit to the length of a polypeptide chain but those occurring naturally tend to be less than 1000 residues. This may represent a constraint derived from the fidelity of translation in the synthesis of the protein (or a historical relic from the days when fidelity was poorer) or it may simply be a consequence of the time needed to synthesise the protein. There are, of course, many exceptions and the largest known protein has about 100,000 residues (Higgins *et al.*, 1994). Clearly, to fold such a protein into a unique structure would be a formidable task and proteins of this size are composed of repeated units: either of like or mixed type. When the repetition is regular, involving a single (or few) type(s) then the resulting structure takes the form of a general helix — providing there is good interaction between the repeats. Otherwise, if the repeats form independent units, the structure has the form of a flexible string of beads. These proteins are referred to as fibrous and tend to play a more inert structural role in the cellular functions.

2.1.2 Globular proteins

Of greater interest are the proteins that have a unique structure derived from a non-repetitive sequence. These tend to fold in to fairly compact units and are, correspondingly, referred to as globular proteins. This class is composed predominantly of proteins in the size range of a hundred to several hundred residues. They include the majority of proteins that catalyse metabolic processes (enzymes) and those that regulate replication and expression of the genetic material. Clearly this covers most of the interesting functions of life and this richness is reflected in a corresponding richness of structure. Fortunately, this class is also that about which most is known structurally. This is a consequence of the ability of many globular proteins to crystallise and hence have their structure determined by X-ray crystallography. For the smaller members of the family, the technique of Nuclear Magnetic Resonance (NMR) is also yielding an increasing number of structures.

2.1.3 Membrane proteins

A third class of proteins is restricted to the unique environment of the phospholipid bilayer membrane that surrounds all cells and many sub-cellular organelles. These proteins cover a range from globular proteins that happen to have a small tail that anchors them to the membrane through proteins that are half-in/half-out of the membrane, to proteins that are fully embedded in the membrane. In function, they cover the transport of material across the enclosing cell membrane, ranging from simple ions to the import of nutrients and the export of products that can influence the surrounding environment. For multicellular organisms, one aspect of the latter function is to influence the state or behaviour of neighbouring cells. This can be effected through the secretion of chemicals that others detect

(and, again, the detection involves membrane bound proteins called receptors), or through direct physical contact between receptors.

2.2 The hydrophobic core

Globular proteins generally exist in the aqueous ('soup'-like) environment of the cellular cytoplasm. The basic organising principle of their structure is to get the amino acid side-chains that are not soluble in water (referred to as hydrophobic) together in a core and surround them with a shell of water-soluble amino acid side-chains (referred to as hydrophilic or polar) which provide an interface to the solvent (Figure 2). This arrangement generally results in a protein that is itself soluble in water and prevents unspecific protein-protein aggregation as might occur if the 'sticky' hydrophobic residues were exposed.

2.3 Secondary structure

One complication of this simple scheme, however, is that all residues also have polar atoms in their main-chain and this includes the hydrophobic residues which we would otherwise like to see buried in the core. Burying these residues will now necessarily entail the burial of a polar amide (N-H) and carbonyl (C=O) group with each residue (each of which carry a partial charge).

A solution to this problem is to form a hydrogen-bond between these unlike charges using groups from different parts of the main-chain. When mutually satisfied in this way, the bonded pair can then be 'safely' buried away from solvent. One might imagine that such a pairing could be achieved in an *ad-hoc* manner (simply matching-up whatever pairs came nearby) — but possibly as a consequence of the complexity of connecting such a network, the hydrogen-bonded networks found in proteins are remarkably regular.

Hydrogen bonded pairings are dominated by the shortest local connection along the chain that can be made without significant distortion of the bond geometry — bonding the carbonyl group of residue i to the amide group of residue $i + 4$. When repeated along the chain, this arrangement is a helical structure of period 3.6 residues, known as the α -helix. The second, and almost only other solution of structural importance in proteins (known as β structure), is formed by two remote parts of the chain lining-up to form a 'ladder' of hydrogen-bonds between them. This 'ladder' of bonds can be formed either when the juxtaposed chains run parallel or antiparallel. Each β -strand can contribute to two ladders, allowing the hydrogen-bonded network to extend indefinitely in either direction, resulting in a general sheet structure, referred to as a β -sheet.

Together the α -helix and β -sheet structures are referred to as **secondary** structure, being intermediate in a structural hierarchy in which the polypeptide chain is **primary** and the folded chain is **tertiary** (Crippen, 1978). However there is a wide variety of other commonly occurring sub-structures that cannot

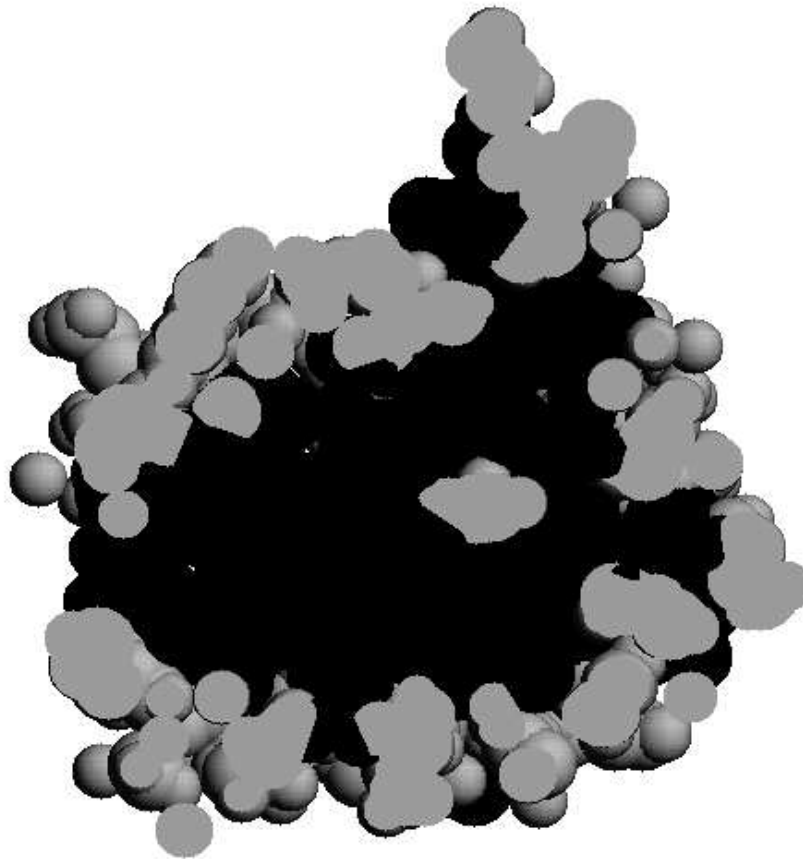


Figure 2: **The hydrophobic core** A section (slab) has been taken through the core of a small protein (PDB code: 3chy) and displayed (using RASMOL) to show the van der Waal's surface of all the (non-hydrogen) atoms. These are coloured as grey for polar amino acids and black for hydrophobic amino acids. The black hydrophobic core can be clearly seen but (as with all 'rules' concerning protein structure) there are some exceptions and a (grey) hydrophilic residue can be seen in the core and a (black) hydrophobic residue on the surface. The former probably is hydrogen bonded to another hydrophilic side-chain or to main-chain polar groups, while the latter may make contact with another protein.

be ignored in a more detailed analysis (Efimov, 1993) including recurring combinations of secondary structures (Efimov, 1991a; Efimov, 1991b; Efimov, 1987) commonly referred to as **isuper-secondary** structure.

2.4 Packed layers

The simplicity of having effectively only two secondary structures is that there are only three (pairwise) combinations of them that can be used to construct proteins; so giving the three major structural classes: 1) α with α , 2) α with β and 3) β with β (Levitt and Chothia, 1976). (For detailed analysis of each class, see: 1) Chothia *et al.* (1981), Lesk and Chothia (1980); 2) Cohen *et al.* (1981), Chothia and Janin (1981), Chothia and Janin (1982), and 3) Cohen *et al.* (1982). With the main-chain atoms tied-up in secondary structure, a core can be constructed using any mixture of α or β building blocks. Incorporation of a β -sheet, however, imposes a long-range constraint across the structure. The β -sheet has free hydrogen-bonds on its two edges, which consequently prevents the sheet from terminating in the hydrophobic core. This divides the core into two and, if considered more generally, imposes a layered structure onto the further arrangement of secondary structures in the protein. (See Figure 3 for examples). (See both Chothia and Finkelstein (1990) and Finkelstein and Ptitsyn (1987) for further consideration of protein structure along these lines.)

2.4.1 All- α proteins

The all- α protein class is dominated by small folds, many of which form a simple bundle with helices running up then down (Figure 3(b)). The interactions between helices are not discrete (in the way that hydrogen bonds in a β -sheet are either there or not) which makes their classification more difficult (Lesk and Chothia, 1980). Set against this, however, the size of the α -helix (which is generally larger than a β -strand) gives more interatomic contacts with its neighbours (relative to the a β -strand) allowing interactions to be more clearly defined. (Figure 3(b)).

2.4.2 All- β proteins

The all- β proteins are often characterised by the number of β -sheets in the structure and the number and direction of β -strands in the sheet. This leads to a fairly rigid classification scheme (Richardson, 1977) which can be sensitive to the exact definition of hydrogen-bonds and β -strands. Being less rigid than an α -helix, the β -sheets can be relatively distorted — often with differing degrees of twist of fragmented or extra strands on the edges of the sheet. (Figure 3(a)). Various patterns can be identified in the arrangement of the β -strands, often giving rise to the identification of recurring motifs (Hutchinson and Thornton, 1993).

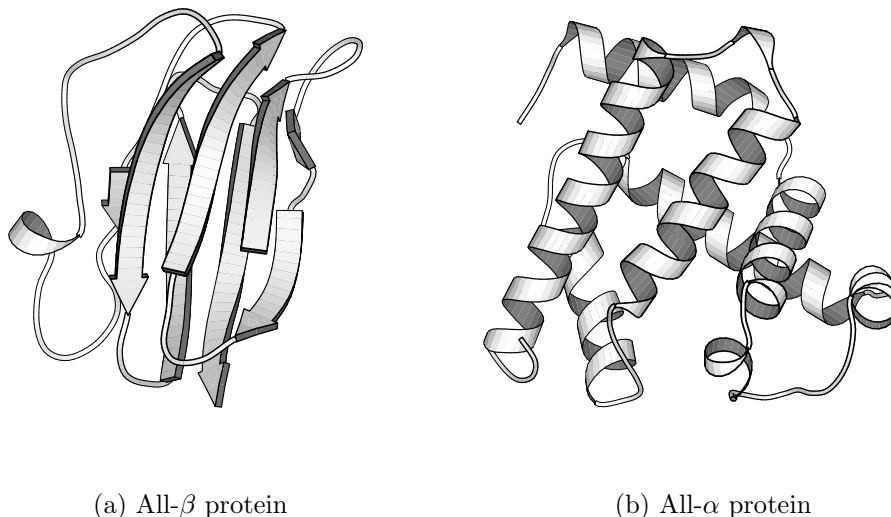


Figure 3: **Protein structures with one secondary structure type** (a) an all- β protein (immunoglobulin) with two packed β -sheets. (b) an all- α protein (globin) showing packed α -helices.

2.4.3 α - β proteins

The α - β protein class can be subdivided roughly into proteins that exhibit a mainly alternating arrangement of α -helices and β -strands along the sequence and those that have more segregated secondary structures. The former class includes structures in which the secondary structures are arranged in layers and those that form a circular or barrel-like arrangement. (Figure 4). Recurring folds can also be identified in the latter type (Orengo and Thornton, 1993).

2.5 Barrel structures and β -helices

Solutions can be found to tie-up the ‘loose’ hydrogen-bonds on the edge of a β -sheet. One commonly encountered, is to twist the sheet so that the two edges meet and can hydrogen-bond to each other — forming a closed barrel-like network of hydrogen-bonds (Chou *et al.*, 1990). This cannot easily be accomplished with less than six strands and if only β -structure is used, then the barrel must incorporate antiparallel pairings. However, in combination with α -helices it is possible to link one (open) end of the barrel to the other and allow the formation of a, predominantly, or pure parallel sheet. A particularly striking example of this arrangement is seen in the eight-fold β - α -barrel $(\beta\alpha)_8$ which was found originally in the enzyme triosephosphate isomerase and is often referred to as the TIM-barrel (Figure 5). (See Murzin *et al.* (1994a) and Murzin *et al.* (1994b) for a full

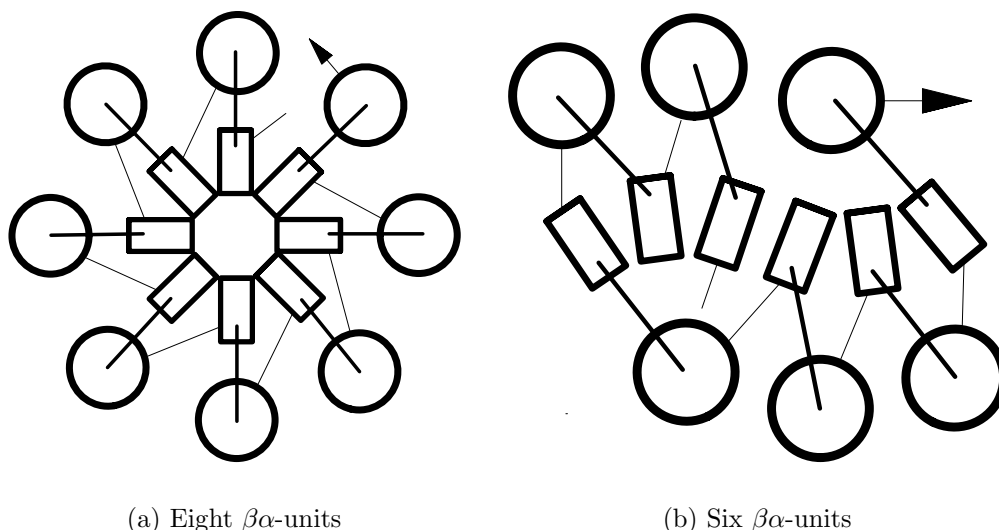


Figure 4: **Folding options for tandem $\beta - \alpha$ units.** β -strands are represented by rectangles and α -helices are represented as bold circles. All strands run parallel and progress towards the viewer. In reality, the strands are both curved and twisted which is suggested by their non-linear alignment and α -helices are about twice as broad as a β -strand. The direction of the chain is indicated by a terminal arrow-head. The structural implications of the size difference between the two secondary structure types (combined with chirality constraints on their connection) are shown for a concatenation of both eight and six $\beta - \alpha$ units. (a) With eight units the sheet can form a barrel and the different radii of this circular form at the β and α level accommodate their different size. The barrel structure is found in many (unrelated) enzymes, typified by triosephosphate isomerase (TIM) and is referred to as a *TIM-barrel*. (See Figure 5). (b) Six units cannot form a barrel forcing an inversion in one half of the sheet to allow helices to be placed both above and below. The resulting arrangement has two-fold symmetry and occurs widely among di-nucleotide binding proteins. It is typified by the dehydrogenases where it is referred to as a *Rossmann fold*.

analysis.)

A barrel can also be formed with the β -strands running in the orthogonal direction (leaving free hydrogen-bonds on the open ends of the barrel). This structure, however, completely dictates the course of the protein chain (as a simple helix) giving little scope for evolutionary exploitation of the fold for different functions. (See Chothia and Murzin (1993) for some examples). This type of structure is associated more with structural (fibrous) proteins.

2.6 Protein Topology

The path of the chain through the various layers of packed secondary structures described above (sometimes referred to as frameworks or architectures) is referred to as the fold of the chain. As this entails various degrees of cross-linking through hydrogen-bonds, it is also possible to, loosely, view it from a topological perspective. This topic will be returned to in detail in Part III but, firstly, a few basic aspects will be considered here which are relevant to the later discussions. (See Ptitsyn and Finkelstein (1980) for a general review.) The course of the chain through the secondary structure frameworks is largely unrestricted. Two constraints, however, are well observed. The strongest is that two loops cannot cross on the same face between layers (Ptitsyn and Finkelstein, 1980)². The source of this constraint is a simple consequence of the bulk of the polypeptide chain: if two loops cross, one will be buried by the other which will be energetically unfavourable unless the buried loop can satisfy its main-chain hydrogen-bonds. Having done this however, the loop is now probably a secondary structure and so the rule that loops do not cross is preserved.

The second strong constraint derives from the chiral nature of the central (α) carbon in each residue. This favours a particular (right) handedness for the α -helix and a corresponding twist to the β -sheet which is left-handed when viewed along the chain direction. Together, these local chiralities result in a strong preference for connections between strands in the same sheet to be right-handed (even when there is no α -helix involved). The few exceptions to this rule are seen when the chain meanders to a remote part of the structure (another domain) and the ‘context’ of the local constraint is lost (Sternberg and Thornton, 1977b) (Figure 6). Some chiral effect are also detected in the $\beta\beta\alpha$ and $\alpha\beta\beta$ arrangements (Kajava, 1992) and in the packing of four α -helices (Weber and Salemme, 1980; Presnell and Cohen, 1989).

2.7 Domain structure

Large hydrophobic cores are not found in globular proteins, probably because of limitations in the folding kinetics and stability. Single compact units of more

²An exception has been found in the protein with PDB code 2csmA.

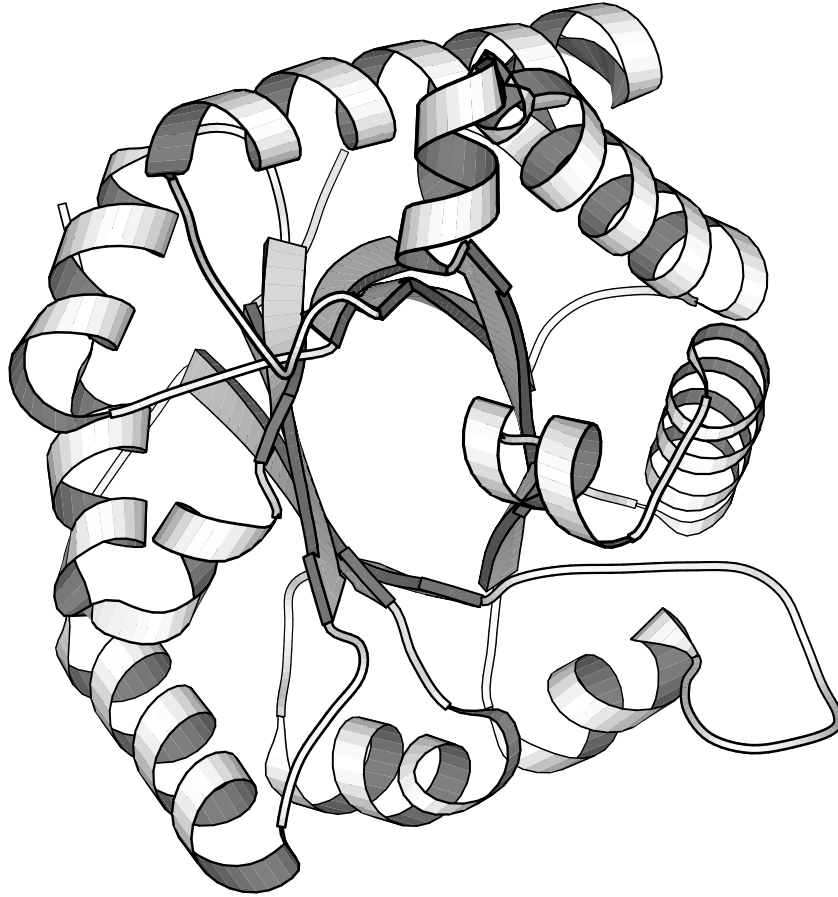


Figure 5: **Eight-fold alternating β/α barrel protein** The protein chain spirals (as a toroid) while alternating between β and α secondary structure type, giving rise to a closed ring or barrel β -sheet in the centre surrounded by a larger ring of α -helices on the outside. The structure, first seen in the enzyme triosephosphate-isomerase (after which it is often named as the TIM-barrel) has been seen many times in unrelated proteins.

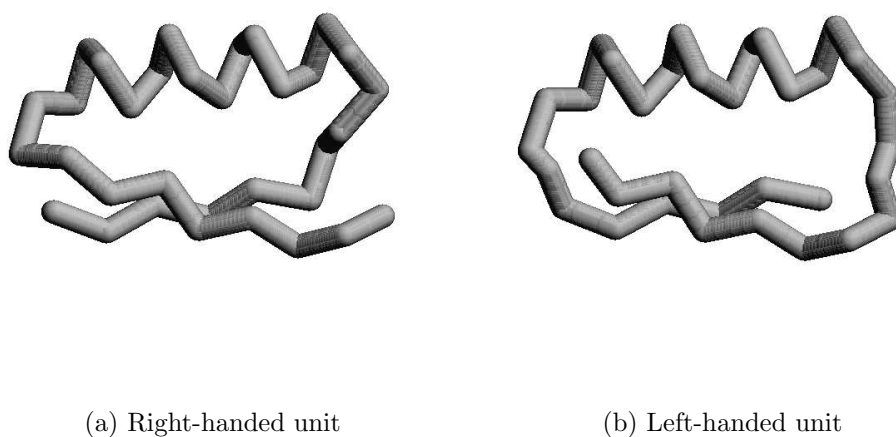


Figure 6: **Handedness in secondary structure connections.** An α -helix linking two β -strands (hydrogen bonded in a sheet) is shown as a backbone (alpha-carbon) trace in: (a) the common right-handed configuration, and (b) with the rare left handed connection. The different chiralities can be appreciated if the whole chain is viewed as a super-helix: in the R-hand form clockwise rotation would drive it into the page (like a screw or cork-screw) while the same rotation would extract the L-hand form.

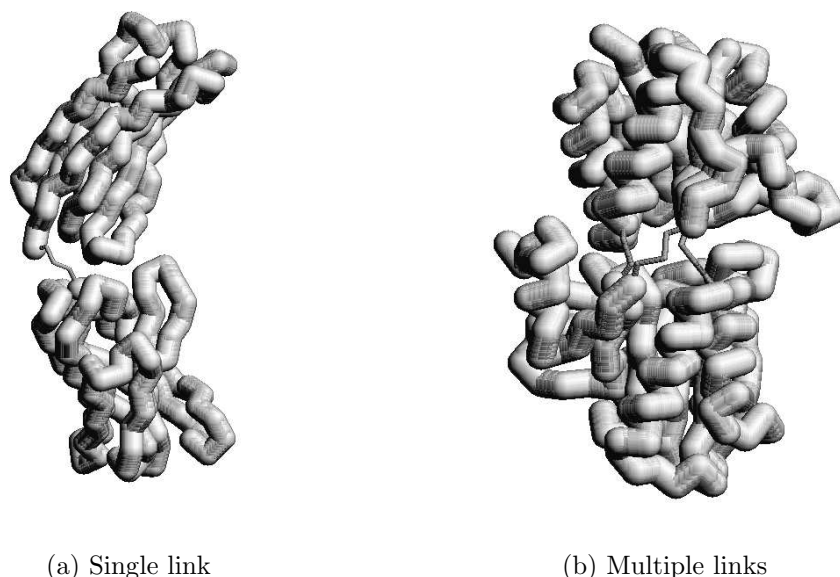


Figure 7: **Simple and complex domain connections.** (a) two immunoglobulin domains linked by a single connection. (b) two more closely packed domains (arabinose-binding protein) between which the chain passes three times. (The linkers have been drawn thinner for clarity).

than 500 residues are rare with the typical size lying more around half this size (200–300 residues). As a consequence, large proteins are organised into units of this size referred to as domains (Rose, 1979; Richardson, 1981; Janin and Chothia, 1985) (Figure 7).

The definition of a domain is problematic — one suggestion is that, if the chain were to be cut, then the two parts would remain stable (with each having its own hydrophobic core). With well segregated domains (like beads on a string) this is undoubtedly true but with more closely interacting domains (and in particular, those in which the chain crosses between the domains more than once), such an experiment cannot be carried-out without exposing surfaces that are not optimally evolved for solvation.

Various working definitions of domains have been derived (Holm and Sander, 1994a; Swindells, 1995; Siddiqui and Barton, 1995; Islam *et al.*, 1995; Sowdhamini and Blundell, 1995) but in the more difficult examples, these seldom agree. The problem with all these methods is that they try to imitate (human) expert definitions and it is clear that these definitions entail the synthesis of many abstract ideas such as biological function, recurrence and symmetry, all of which are difficult to capture in an automatic method. A recent approach to this problem has been based (loosely) on an Ising model, in which structural domains evolve in competition with each other for residues in the protein (Taylor, 1999c).

Part II

Protein Structure Comparison and Classification

Perhaps the most remarkable features of the [myoglobin] molecule are its complexity and lack of symmetry. The arrangement seems to be almost totally lacking in the kinds of regularities which one instinctively anticipates, and it is more complicated than any theory of protein structure.

John Kendrew *et al.* (1958)

3 Overview of Comparison Methods

3.1 Challenges for Structure Comparison Methods

The vast variety of protein sequence and structure found in the current databases could not have been anticipated by a polymer chemist looking only at bonds and forces. Indeed, the best effort from pure stereochemical considerations was made by Linus Pauling who predicted the α -helix from first principles before any protein structures were known. However, this did not prepare people for the sight of the first structures, which were much more irregular than expected³ The most important things we know about proteins have come therefore, not from theory, but from observation and the comparison of sequences and structures.

Equivalent proteins from related species usually have similar structures and sequences and a comparative analysis of these can tell us about residue substitutions and how the structure adapts to accommodate them. However, if one is interested in the stability and versatility of protein structure under greater degrees of sequence variation — in other words, how far can a structure be pushed by evolution, then it is necessary to compare the most distantly related proteins. This has driven those who develop methods to compare protein structures to continually ‘push-back’ the range of comparison methods with the hope of discovering further and perhaps more fundamental similarities among proteins.

Another way of viewing this problem is to consider an abstract space of all sequences. As we have seen, one sequence gives rise to one unique structure. However, the mapping in the other direction is not unique and many sequences can give rise to the same structure. We can then ask, how big is the space of sequences that give rise to the same structure. The answer to this will tell us how stable different folds are and whether through, ‘random’ evolution some are more likely to accumulate than others. These fundamental questions raise further questions: most importantly, “at what degree of dissimilarity do we consider two proteins to be the same or different?”. Without an answer (or an approach to) this problem we cannot get answers to the other more fundamental questions.

The same problem is encountered from an evolutionary angle: if our prime interest is in determining whether two proteins are evolutionarily related (share a common ancestor). Again, such questions cannot easily be answered as the probabilities depend on how accessible a common fold is to different sequences. In other words, we need the complete map of sequence space — annotated by structure. With many assumptions, this might be roughly estimated if every distinct type of protein structure (topology or fold) were known along with their frequency of occurrence. However, given our limited data, the question of “how many folds?” is not easily answered and begs the question of whether those we see in Nature are a complete covering of the possibilities or represent a fraction

³See the opening quote to this Part.

that has been ‘frozen’ through some historical accident. (These problems will be returned to in Section 8).

To tackle questions such as these, methods of protein structure comparison have been developed. These have been based on a great variety of approaches aimed at different aspects of structure (folds, fragments, etc.). With improved computer power, some of these methods have even been applied to the comparison of the complete protein structure databank, giving an automated analysis of what had previously been the monopoly of a few experts. To assess the results of such analyses (which will be considered in Section 8) it is necessary to know how the various methods work as many behave quite differently. Such knowledge is equally vital when choosing a method for a particular comparison problem.

3.2 Degrees of Difficulty

Structural similarity is of interest at many levels, from the fine detail of backbone and side-chain conformation at the residue level, through the coarse similarity of folds at the tertiary structure level, to a simple count of secondary structures. Similarities may also be locally confined or extend globally over whole structural domains and even involve more than two structures. These issues are reflected in the methods that will be discussed below: spanning comparisons of almost identical structures through to highly dissimilar ones.

The simplest applications are concerned with studies on a single protein. Examples include studies of conformational change between states of the same protein (including multiple NMR structure solutions), and the comparison of mutant forms of a protein where the structures being compared usually have very similar structure at all levels of detail and negligible or no insertions and deletions of sequence (indels).

Applications of intermediate difficulty include comparison of closely related proteins to analyse evolutionary divergence, inference of weak sequence homologies on structural grounds, characterization of conserved structural features such as functional sites within families. Conversely, structure comparison may help in the analysis of similar folds that apparently result from evolutionary convergence (Orengo *et al.*, 1993). Sometimes the requirement is to screen a specified structural fragment (motif) against a database of protein structures, searching for strong matches. In these examples, the structures of interest are relatively similar, so that indels present a limited problem.

The most difficult and general structure comparison applications arise in the classification of the known protein structures into different fold families. This rationalizes the organization of the structure databank, and may indicate hitherto unsuspected structural similarities, evolutionary relationships, or constraints on folding. Powerful comparison methods must be able to deal with structural similarity at all levels of detail, must handle indels of arbitrary length and position in the respective structures, and must identify structural similarities even when

these form a relatively small proportion of the structures being compared.

This diversity of applications is addressed by a corresponding variety of automatic or semi-automatic comparison methods, some suitable for comparing highly similar structures at a specific level of detail or *element size* (residue, backbone fragment, secondary structure, etc.), while other more general methods may operate at several element sizes or may be applicable to more remote comparisons.

The common aims of each method are to compute some quantitative measure of similarity, and often to generate a structurally derived alignment of one protein sequence against the other(s). The set of element equivalences so defined may be used to drive a rigid body superposition to facilitate visual comparison, either as an intrinsic part of the method, or as a separate step (Rippmann and Taylor, 1991).

3.3 Different Approaches

Structure comparison methods differ in many ways: these include the basic choice of algorithm(s) and the kind or size of structural elements compared. Finer distinctions are found in tolerance to indels, ability to detect mirror images, translocations, or rigid internal rotations, and (as in sequence comparison) ability to distinguish between local or global similarity. The following sections summarise some of these aspects which are used below to assess the different methods (and what they might be good for).

3.3.1 Comparison Power

The simplest methods rely only on the (bulk) structural content of the protein (such as number of secondary structures). Judged by their ability to distinguish proteins of differing degrees of similarity, these methods can be classed as **weak**. They are generally statistical in approach, thereby losing both the individual identity and the ordering of the component elements. Methods that are **intermediate** in power, are those that preserve the identity of elements and finally there are **strong** methods that preserve both element identity and sequential order. This latter distinction corresponds to that of Rossmann and Argos (1977) who defined the terms *structural equivalence* to describe the spatial similarity of components and *topological equivalence* for connected runs of structurally equivalent components, sometimes referred to below as non-sequential and sequential categories, respectively.

3.3.2 Feature or Relationship

Methods differ also in the type of data structures that they compare. These fall into two classes depending on their definition for a given element. A *feature* is an intrinsic property of each element: this can be a single scalar property value

or a (fixed length) vector of such values associated with an element. Examples for residues might be solvent accessibility (scalar) and $\{\phi, \psi\}$ main-chain torsion angle pairs (vector). By contrast, a *relationship* describes each element in terms of other elements in the structure: for each element there is a relationship value with every other element, an example being interatomic distance. The essential difference between a feature and a relationship set is that, for any protein with N structural elements, the number of feature values (or compound feature values, e.g., $\{\phi, \psi\}$ angles) is proportional to N , while the number of relationship values is proportional to N^2 .

The simplest comparison approach might be to define a measure based only on features: say, the secondary structure state and degree of burial of the two residues in the two proteins being compared. Such a simplistic measure, however, could not distinguish two adjacent β -strands both of which were buried in the core of both proteins. For this, a description of environment is required that can capture the true 3-dimensional relationship between residues (their topological relationship). This poses a difficult computational problem and might best be appreciated by the following simple example. Consider two β -strands — A and B, found in both proteins being compared and lying in the order A-B, both in the sequence of the two proteins and also in their respective β -sheets. If both pack against an α -helix then, in both proteins, a point on A would be buried by a β -strand to the right and an α -helix above, and would be considered to be in similar environments. If, however, in one protein, the α -helix lay between strand A and B, while in the other protein it lay after strand B then the two arrangements would not be topologically equivalent (Figure 8).

3.3.3 Hybrid methods

Some methods operate with more than one element size and/or structural property and function as discrete multi-stage or combined algorithms (which are sometimes iterated). Many of the most recent developments behave in this way and these are best described as hybrid methods but can generally be decomposed into their components using the ideas described above.

3.4 Dynamic Programming

3.4.1 The basic evolutionary model

The ability of proteins to lose or gain sequence elements over evolutionary time (relative insertions or deletions: jointly referred to as *indels*) has led many methods of structure comparison to follow the simple model of evolutionary change which is used in sequence alignment methods. This assumes that the only processes at work are substitution of amino acids (or rather the underlying nucleotides) and their deletion or insertion. More complex operations such as re-

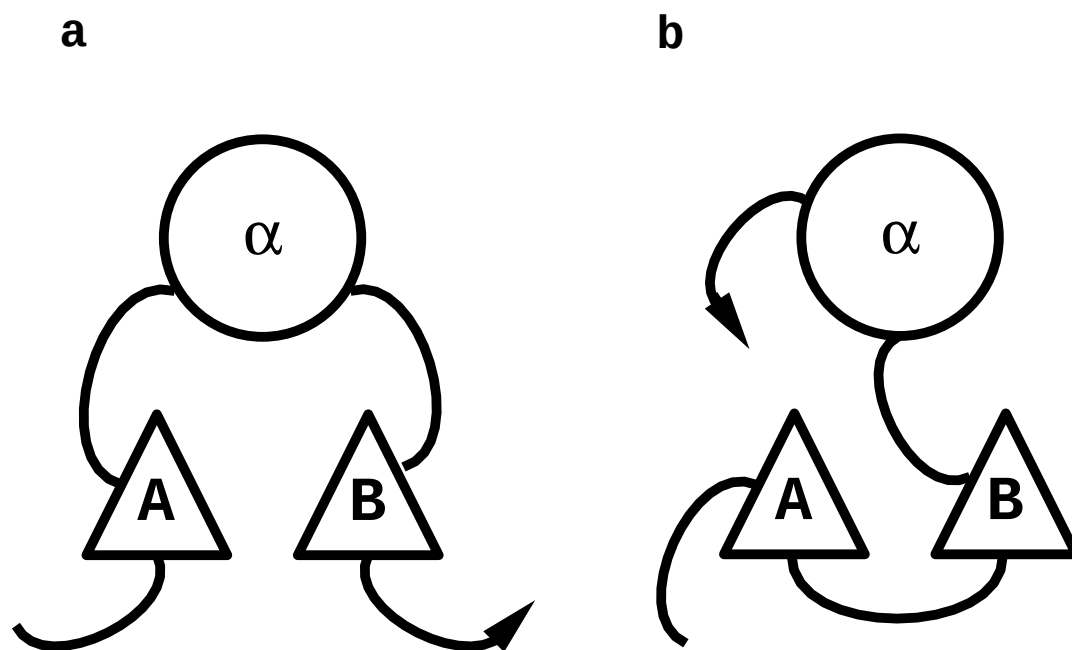


Figure 8: **Topological and structural equivalence.** Two β -strands, A and B, are shown schematically as triangles packing against an α -helix(circle) in two distinct structural fragments, **a** ($\beta\alpha\beta$) and **b** ($\beta\beta\alpha$). The packing in the two fragments could be identical but a comparison method that takes account of the topology (or connectivity) of the units would not detect any great similarity.

versals, translocation and duplication events are ‘forbidden’. This model further assumes that these processes are uniformly applied along the sequence length and are the same for all proteins. In addition, most alignment methods implicitly assume that the substitutions⁴ in one place do not affect substitutions elsewhere. From our knowledge of protein structure this latter assumption is clearly untrue (one part of a structure can influence any other part) but, despite this, the sequence alignment model provides a good starting point.

This model of sequence evolution is implemented in a simple algorithm called *dynamic programming*. As this algorithm will be referred to frequently below, it will be described here — initially in the context of sequence comparison followed by an outline application to structural data.

3.4.2 Sequence Alignment

The alignment of one sequence with another can be represented by constructing a grid (or matrix) with a sequence on each axis. Each cell in this matrix links a pair of elements (residues or nucleotides) in the two sequences and an alignment of the two sequences is a path through the matrix that progresses without any backwards or stationary steps in either sequence. The problem to be solved is to find the path through the matrix (top or left edge to the bottom or right edge) that passes through the highest scoring cells finding a maximum sum of scores. (See Figure 9 for a worked example). This algorithm is guaranteed to find the optimal alignment under a given scoring scheme, providing pairwise matches are independent: that is; the score for each match is unaffected by matches elsewhere. (See Pearson and Miller (1992) for a review).

The dynamic programming algorithm forms the basis of many widely used sequence alignment algorithms which can align one whole sequence against another, giving an overall or *global* alignment (Needleman and Wunsch, 1970), or find the section that aligns best, giving a *local* alignment (Smith and Waterman, 1981). In general, any information that can be encoded as a sequence (providing the elements can be matched independently) can be aligned using the basic dynamic programming algorithm. For proteins, this can be either pure sequences or can be structural data (encoded as a string) allowing ‘structures’ to be compared against each other and with an amino acid sequence.

3.4.3 Gap-penalty

A penalty against gaps in a sequence alignment can easily be incorporated into the standard dynamic programming algorithm simply by subtracting a constant value from each score inherited by any transition other than the diagonally adjacent cell. This can impose a fixed penalty for any size of gap but can be refined to be partly (or wholly) dependent on the gap size. (See Figure 9 for examples.)

⁴Substitutions are realised as mismatches in an alignment of two sequences.

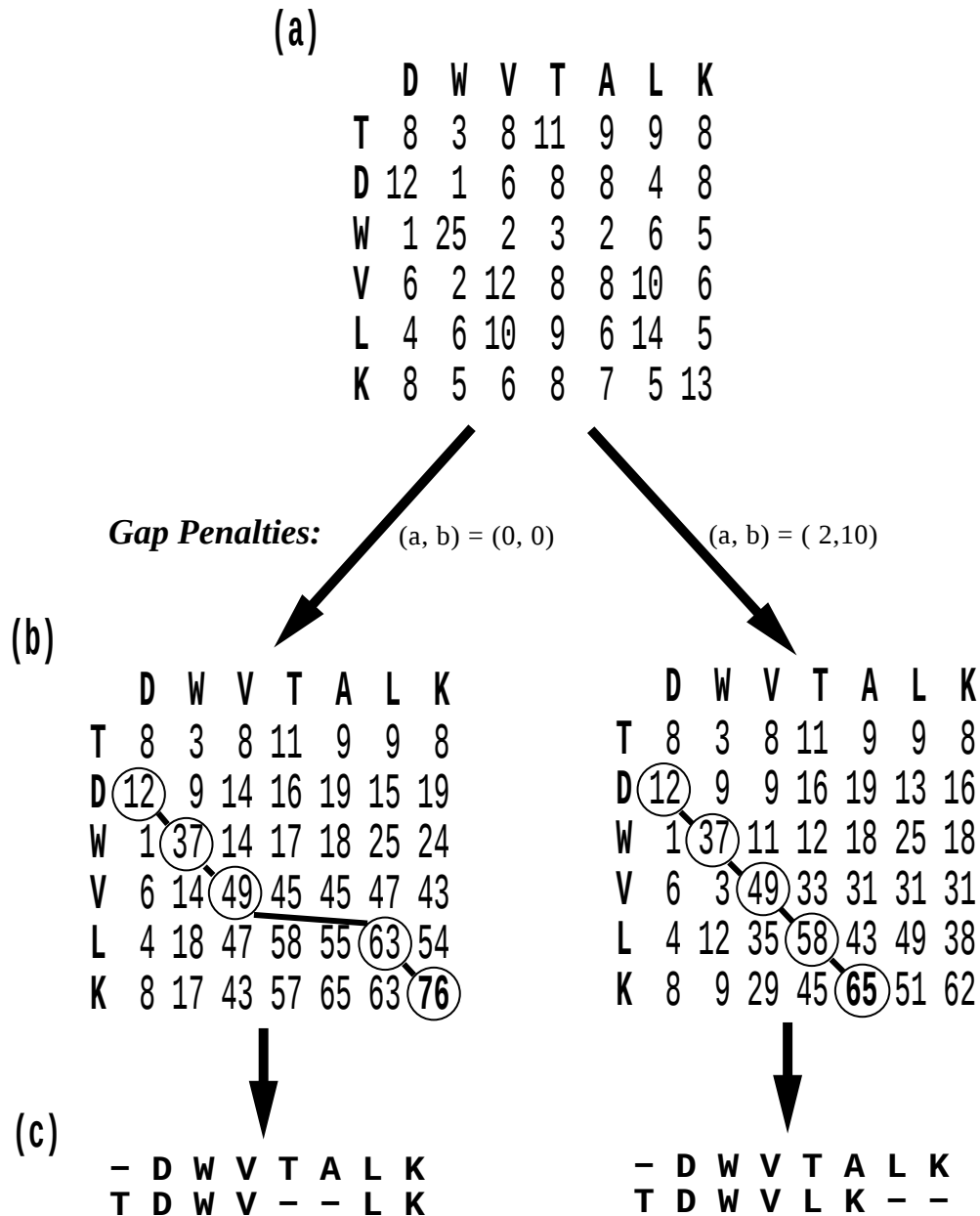


Figure 9: **The basic Dynamic Programming algorithm.** Steps in this algorithm are illustrated using the alignment of two short protein sequences. (a) the sequences form a matrix in which each element is a similarity score for the match. (Different pairs get a different score: *e.g.* a W:W match gets 25). (b) the matrices are summed with each cell (i, j) adding the best summed score in the previous sub-matrix (all with indices $< i, < j$) to its own. (*e.g.* W:W = $12 + 25 = 37$). (c) Two variations are shown one in which gaps are free and the other where there is a high cost resulting in an ungapped alignment.

The form of the gap penalty reflects an underlying model of protein evolution which, in part, is a reflection of the stability of protein structure. A simple model for the latter is that the core is very sensitive to change while the surface, and especially exposed loops, are very susceptible to change (Bajaj and Blundell, 1984). Indeed, it probably makes little difference to a protein structure whether 10 or 100 residues have been inserted into an exposed loop — providing the insert is in the form of a compact, independently folded, domain. On this basis, the simple gap-penalty (no extension penalty) provides an adequate model for remotely related proteins.

When the more complex form for the penalty is employed there is one penalty to open the gap and another making it dependent on gap size. A linear function with positive coefficients is commonly used: $an + b$, where n is the gap length ($a = 1, b = 10$; are typical). This is generally referred to as an *affine* model (Altschul and Erickson, 1986).

A further general result from the analysis of affine gap-penalties emerges in the phase-space of the two gap parameters (a and b) where the alignments are found to fall into two types: 1) when the penalty for a gap is high the best ungapped (or local) alignment is optimal and, 2) when the penalties are reduced, a boundary is crossed (a phase transition) into the region of gapped alignments. Correct protein sequence alignments seem to lie close to this boundary (Vingron and Waterman, 1994).

3.4.4 Structure Biased Gap-penalties

If the three-dimensional structure of one (or both) of the proteins is known then the model of what gaps might be possible at different locations on the sequence can deviate greatly from the simple problem of aligning two sequences. Relative insertions and deletions of sequence are much less likely to be found in segments of secondary structure segments, especially when these are buried (Pascarella and Argos, 1992; Johnson *et al.*, 1996). If the structure of the proteins are known, then the alignment program can use this information and modify its local gap-penalty accordingly to avoid breaking secondary structures and inserting residues in the hydrophobic core (Lesk *et al.*, 1986; Barton and Sternberg, 1987; Kanaoka *et al.*, 1989; Zhu *et al.*, 1992; Smith and Smith, 1992; Johnson *et al.*, 1996).

4 Early and Simple Approaches

Most of the early methods of structure comparison were developed by crystallographers in order to compare their new structures with others (or themselves, if they contained internal duplications). Mainly, they depend upon transformation of the global coordinate frame of one molecule into that of the other and therefore

tend to be most successful when comparing closely related structures. These algorithms rely on minimising the root-mean-square deviation between equivalenced positions (McLachlan, 1972a; Kabsch, 1976; Sippl, 1982).

Parallel to these developments of simple superposition, a different line developed that was based on comparing structural features that had been reduced to strings (sometimes using the dynamic programming method described above). These two approaches prepared the ground for the development of later methods (described in the following sections) that synthesised the true 3D structural view with an alignment model.

4.0.5 Manual and semi-automatic methods

Many of the older applications of rigid body superposition rely on manual specification of the topologically equivalent residues, for example Muirhead *et al.* (1967); McLachlan (1979); Schulz (1980). Useful as such methods are for specific comparisons between closely related structures, some means of semi-automatic or automatic selection of equivalences is necessary if large numbers of comparisons are to be performed, or if detailed knowledge of topological equivalences is lacking.

Rao and Rossmann (1973) and Rossmann and Argos (1975) were the first to describe a semi-automatic iterative method which was manually primed with a set of topologically equivalent residue pairs. The two molecules are superposed using a least-squares procedure which searches Eulerian angle and vector space to minimize the RMS (root-mean-square) score between equivalent residues. Given the new spatial correspondance, probabilities relating to spatial similarity and orientation are computed for each residue pair between the two proteins. Those sequential pairs having the highest probabilities form a new equivalence set, which are then used to drive another transformation. Repeated cycles of equivalence assignment and transformation are applied until the equivalence list is stable. The original method requires prior knowledge of equivalences but Rossmann and Argos (1976) eliminated the manual priming by employing a search function in rotational space. This attempts to maximize the number of equivalences while three rotational axes are systematically explored.

The Rossmann-Argos method was able to determine the positions of limited indels as long as sequential sets of equivalent residues can be identified, but, by comparison with the contemporary Remington-Matthews procedure described below, is computationally demanding, and requires careful tuning (Matthews *et al.*, 1981; Matthews and Rossmann, 1985).

4.0.6 Fragment based methods

In the method of Remington and Matthews (1978,1980) all possible backbone segments of a given length in the first protein were compared with those of the

second protein by a rigid body superposition. The resulting distribution of RMS scores was then used to assess the statistical significance of high scoring segment pairs. Additionally, an RMS contour map of one protein sequence against the other reveals these pairs as peaks in relation to the two sequences. The transformations applied to segment pairs contributing to any dominant peak are an estimate of the transformation for overall superposition. The method is analogous to the similarity (dot-plot) matrix used in sequence comparison and is useful for identifying structural repeats.

Similar segments in separate proteins can be located provided they are not interrupted by large insertions and deletions. In contrast to the contemporary Rossmann-Argos method described above, the Remington-Matthews method is easier to apply, computationally less demanding, and yields a statistical significance for any superposition (Matthews *et al.*, 1981; Matthews and Rossmann, 1985).

The fragment-based dynamic programming method of Zuker and Somorjai (1989) defines a distance measure based on rigid body superposition of C α backbone fragments of three or more residues in one protein onto their counterparts in the second protein. Dynamic programming was used to identify a set of maximal length, non-overlapping aligned fragments separated by indels, which produce an overall best superposition. A penalty is applied for breaking fragments and this prevents the solution from degenerating into the trivial case of superposing fragments of length three. A large number of superpositions must be performed to determine the best possible set of fragments, and the authors developed their own fast superposition algorithm based on quaternion algebra.

4.1 Comparing Feature Strings

4.1.1 Residue level

Levine *et al.* (1984) use the sequence of backbone torsion angles to compare two proteins, constructing a matrix of the combined difference score for the main-chain torsion angles (ϕ and ψ). The matrix can be analysed in a number of ways to obtain an overall similarity score between the two proteins. These include a fast method of searching for the best path through the matrix using lists, while a second method obtains a statistical measure of similarity based on the distribution of torsion angle similarity values along the diagonals. The statistical nature of this method may lead to the significance of some comparisons being missed. A further problem is that similarities in secondary structure will be identified regardless of topological equivalence and this is liable to complicate the measurement of global similarity.

4.1.2 Backbone-fragment level

In a generalization of the residue torsion angle method of Levine *et al.* (1984) (above), fragments of backbone have been examined by Karpen *et al.* (1989), who compare all possible fragments of a chosen length from one protein with those in the other. Fragments are compared using a RMS measure computed over their torsion angles, this score being recorded in a matrix as in Levine's method. The technique was intended only to identify and rank local features rather than to produce an overall similarity score.

The approach of Rackovsky and Scheraga (1978), (Rackovsky and Scheraga, 1980; Rackovsky and Scheraga, 1984) uses differential geometry to describe the trajectory of the protein backbone approximated as a discretized curve. Each segment in the chain is parameterized by a curvature and torsion computed from the α -carbon coordinates of a tetrapeptide, this being the smallest applicable backbone fragment. However, the method is unsuitable for disparate chains containing indels and within these limitations, it is sensitive to the presence of an internal rotation with respect to similar substructures in the two proteins, as the plots show complete identity except in the region of the rotation.

4.1.3 Secondary structure level

Abagyan and Maiorov (1988, 1989) idealize the protein backbone as a chain of vectors connected head to tail. Vectors alternate along the chain representing linear secondary structure elements (α -helices and β -strands) and intervening loops. The trajectory of the backbone is described by vector lengths, angles between sequential vectors, torsion angles about intermediate vectors, and, in the later work, by the number of residues in each secondary structural element. Their program FASEAR (Abagyan and Maiorov, 1989) combines the four measures to compare structures in a 2D matrix. Runs of minima in the matrix indicate possible topological equivalences, which are used to superpose the vector chains using the algorithm of McLachlan (1979). The method is suitable for crude comparisons of similar tertiary folds or for searching a database for some specified supersecondary motif.

5 3D Methods without dynamic programming

The more automatic methods described in the previous Section simplified each protein to a string, and in so doing lost the essential 3D relationship between features. This can be retained by considering each protein chain as a sequence of elements described by their structural relationships with each other. A simple visual method for examining the internal structural associations of a protein is the distance plot (matrix, map or diagonal plot) due to Phillips (1970), by which a property, typically interatomic distance between α -carbons, is recorded in the

cells of a symmetric 2D matrix, the axes of which are the amino acid sequence of the protein. To compare two proteins A and B (of equal length), it is necessary to construct two such matrices (A against A, and B against B). Being conformable, the matrices can be compared cell by cell and combined in a difference matrix (DM) of the same dimensions, from which an overall difference score may be computed.

Most of the methods described in this section make use of this device but do not use dynamic programming to impose an alignment of the structures.

5.1 Distance-matrix matching

5.1.1 Early attempts

One of the first attempts to compare proteins using this approach was Nishikawa and Ooi (1974b) who derived difference distance plots by subtracting conformable distance matrices representing the two proteins. Equivalenced residues were indicated by a low average score along the row or column and an overall measure of similarity was obtained by calculating the total or average difference score over the whole plot. The latter may be formulated as an RMS estimate of similarity of the two conformations, although problems with mirror images may arise. The technique is less amenable to comparing proteins with larger indels. Dissimilar regions can, however, be excised to make the distance matrices conformable and Padlan and Davies (1975) described a means of stretching a shorter sequence by inserting padding ‘residues’ between known equivalent marker residues in the two chains. A variety of matrix scoring techniques that aim to overcome such problems have been described (Liebman, 1982). These solutions are manual and difficult to apply, so that the basic difference matrix approach is constrained to closely similar conformations.

A variation on comparative distance plots yields a spectrum of difference scores corresponding to successive levels of spatial interaction (Sippl, 1982). These are obtained by comparing subsets of the two distance matrices corresponding to successive off-diagonal elements. The set of scores so obtained contains more information on structural similarity than a single overall measure and local features such as internal rotations of one molecule are identifiable. The different diagonals reveal similarities at different levels of structural organization. For instance, for secondary structures, diagonals of order up to 10 (i.e., distances between the i and $i + 10$ th residue) should be used. Similarly, for domain level organization, orders between 10 and 25 give medium and long range structural information.

5.1.2 The DALI method

Holm and Sander devised a two stage algorithm, implemented as DALI, which uses simulated annealing to build an alignment of equivalent hexapeptide backbone fragments between two proteins Holm *et al.* (1992); Holm and Sander (1993b); Holm and Sander (1993a). The approach is equivalent to aligning collapsed distance matrices of the proteins from which insertions and deletions have been excised — similar to some of the earlier methods described above.

In the first stage, hexapeptide contact maps are matched and similarity scores generated by comparing all distances within the hexapeptides. An ‘elastic’ score proportional to the relative differences between distances is used, making the method more tolerant to distortions in longer range distances. Hexapeptides whose contact maps match above some threshold are stored in lists of fragment equivalences. To reduce the amount of information considered, only hexapeptide pairs having similar backbone conformations are compared. Similarly, although residues occur in a number of overlapping contact maps, the map with the closest contacts to any other segment is selected for a given residue.

In the second stage, an optimization strategy using simulated annealing explores different concatenations of the fragment pairs. Similarity is assessed by comparing all distances between aligned substructures. Each step consists of addition, replacement, or deletion of residue equivalences, in units of hexapeptides and, since hexapeptides can overlap, each step can result in the addition of between one and six residues. In the next iteration step, the alignment is expanded by adding substructures that overlap with those already equivalenced. Once all candidate fragment pairs have been tested, the alignment is processed to remove fragments with negative contributions to the overall similarity score.

An advantage of the approach is that the alignment need not be constrained by fragment sequentiality, so that fragments can be equivalenced in a different order along the sequences. The method has been used to compare representatives from all the non-homologous fold families in the Brookhaven databank (Holm and Sander, 1994b; Holm and Sander, 1997; Holm and Sander, 1998). (see Section 8 for further details).

5.1.3 Backbone fragment methods

The Suppos algorithm incorporated in the WHAT IF program produces superpositions based on either structural or topological equivalence of backbone fragments, and can also permit chain reversal (Vriend and Sander, 1991). The first of three stages identifies similar fragments of a given length (10–15 residues) between the two proteins by rigid body superposition using the algorithm of Kabsch (1978). These are then iteratively grown and superposed until the same threshold is reached, thereby identifying maximal length similar fragments. The final rotation matrix used in this superposition is stored with each fragment. In the second

stage, these pairs form nuclei for a clustering process, in which pairs are fused if their rotation matrices are similar (within some error). The third and final stage checks the similarity of the internal spatial organization of respective clusters in each protein and computes a new superposition based on the largest cluster.

Alexandrov *et al.* (1992) describe an almost identical method implemented in the SARF program. This starts with 6–7 residue length overlapping fragments and superposes using the McLachlan (1979) algorithm, retaining all pair matches better than a threshold in a fragment pool. All neighbouring fragments with similar rotation matrices within a tolerance are united, resuperposed on their counterparts, and the best are returned to the fragment pool. This is in contrast to the pairwise clustering used by Vriend and Sander (1991). Merging and superposition are iterated, sampling from the pool, until the superposition score is stable.

5.2 Secondary structure graph-matching

The comparison of proteins at the secondary structure level developed from some early attempts (Kuntz *et al.*, 1976) through the more complex 'meta-matrix' analysis of Richards and Kundrot (1988) to the automated POSSUM method (Mitchell *et al.*, 1989; Artymiuk *et al.*, 1990) which compares the geometric relationships between α -helices and β -strands abstracted as axial vectors. In this method, proteins are represented as fully connected graphs whose nodes are secondary structure elements and whose edges are pairwise closest approach and midpoint distances and torsion angle. A standard subgraph isomorphism algorithm detects subgraphs in each protein in the database equivalent to that of the query structure, having the same types of nodes with similar valued edges within user specified distance/angle tolerances.

Ordering of secondary structure elements in the query is under user control. There is no alignment score and a detailed residue level alignment is produced by conventional superposition. Like the geometric searching techniques (Lesk, 1979; Brint *et al.*, 1989) to which it is related, the method is unsuitable for the general problem of identifying unspecified common substructure (i.e., discovery of unspecified common subgraphs). Nevertheless, it is appropriate for fast database searching with known motifs to identify candidates for more refined comparison.

The subgraph discovery problem is addressed by PROTEP (Artymiuk *et al.*, 1992b; Artymiuk *et al.*, 1992a; Grindley *et al.*, 1993), which uses an established maximal common subgraph algorithm to compare the same secondary structure types and relationships. This identifies maximal fully connected subgraphs or cliques that are shared between structures. As with POSSUM, the method allows fast database searching, but does not give a residue level alignment or superposition.

Subbarao and Haneef (1991) also represent protein structures as partially connected graphs whose nodes and edges are C α atoms and interatomic distances,

respectively, within some user specified cutoff. Two graphs are compared to identify the maximal common subgraph corresponding to structurally similar regions using a standard algorithm. The set of C α atom equivalences mapped by the subgraph is used to drive a conventional superposition (external feature) from which a new set of C α equivalences within a 3Å limit is used produced to drive a final superposition.

5.3 Geometric-hashing approach

Geometric searching techniques are used in small molecule applications to match a query structure against a database of molecules and Lesk (1979) has described a geometric searching method suitable for proteins or other macromolecules. This computes a sorted list of interatomic distances in the query structure and associates with each atom a bitstring wherein the i th bit is set if the atom has a neighbour at the i th position in the distance list. The bitlist is thus a discrete signature for that atom. Similar bitstrings in terms of the same distance list are constructed for each database structure in turn and compared with those of the query structure to derive tentative atom equivalences. The number of comparisons may be reduced by only considering ‘like’ atoms by some property, e.g., atom type.

The final, and computationally demanding, step samples all combinations of matched atoms for each database structure to find the best equivalence set by superposition onto the query structure. Brint *et al.* (1989), also working with C α interatomic distances, describe an optimization of Lesk’s algorithm, which speeds up the method by replacing the combinatorial sampling step with a backtracking tree search.

Since the query specifies the substructure to be matched, these methods are unsuitable for the general problem of identifying unknown common substructure. In contrast, an application of the computer vision technique termed geometric hashing is suitable for database searches using defined patterns or to discover unknown similarities (Nussinov and Wolfson, 1991; Fischer *et al.*, 1992; Bachar *et al.*, 1993). The method is demonstrated using C α atoms, although any atoms can be discriminated on type, or other properties.

A triple of (non-linear) atoms in a protein defines a reference frame and, in general, all such triangles are computed and the side lengths are hashed to compute an address in a hash table, at which the protein identity and three atom coordinates are stored. The hash table is populated in this way for all proteins. Once compiled, it can be used for any comparison, and new proteins can be inserted without recomputing existing entries. A simplified outline of the algorithm is shown in Figure 10.

Matching a query protein against the database proceeds by looking up the query protein triangles in the hash table. Each match found constitutes a vote for the triangle entries stored at that address. High scoring matches represent

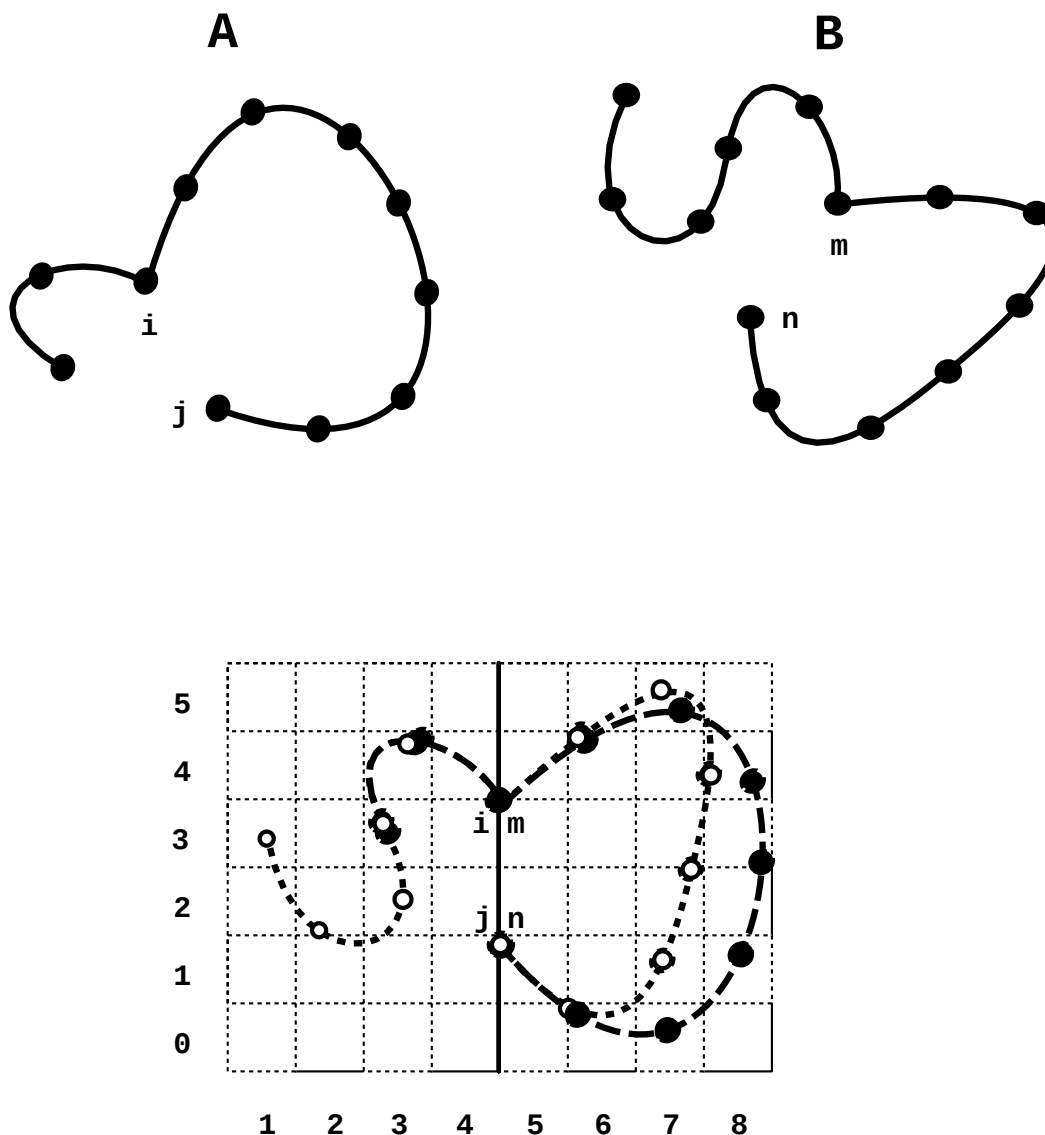


Figure 10: Two protein structures **A** and **B** are shown schematically. Two pairs of positions (i, j in **A** and m, n in **B**) are selected. Both structures are centered on the origin of a grid (below) at i and m and orientated by placing a second atom in each structure (j and n) on the vertical axis which is (coincidentally) the terminal atom of each structure. (In three dimensions, three atoms are required to define a unique orientation.) Atoms in both structures (open and filled circles) are assigned an identifier that is unique to the cell in which they lie (the *hash* key). For simplicity, this is shown as the concatenation of two letters associated with the ordinate with the abscissa (XY). For example; atoms in structure **B** are assigned identifiers AD, BC, CC, CD, etc. The number of common identifiers between the structures provides a score of similarity. In this example these are CD, CE, FE, GF, HE and FA (not counting i, j and m, n) giving a score of 6. The process is repeated for all pairs of pairs, or in 3 dimensions, all triples of triples and the results pooled.

reference frames common to both proteins, and the rigid body transformation required to superpose triangles is an estimate of the overall superposition for the molecule, while the participating atoms are an estimate of the desired atom equivalences. Many matched reference frames correspond to essentially the same transformation, and these are merged to produce a larger set of atom equivalences. These then prime a series of superposition and assignment steps to further refine and extend the equivalence list.

The same basic method has been applied to comparing protein surfaces at ligand binding sites (Fischer *et al.*, 1993; Fischer *et al.*, 1994).

6 3D Methods using Dynamic Programming

6.1 Using structural superposition

Barton and Sternberg (1988) describe a specialized application of dynamic programming to finding residue equivalences. Given an initial superposition based on the cores of two closely related proteins, the LOPAL program determines residue equivalences between variable loop regions, which may differ in the number of residues as well as spatially. Each such region is represented by a distance matrix holding all C α distances from one loop to the other and dynamic programming is used to find the best global alignment, effectively aligning the C α atoms by their 3D coordinates. This approach was later developed into a more general method that used sequence alignment to establish an initial correspondance (Russell and Barton, 1992).

In the later development of this approach (in the program) STAMP (Russell and Barton, 1992), multiple pairwise sequence alignments are used to construct a binary tree ordered by sequence similarity. Structures are then superposed using a conventional pairwise algorithm in the order dictated by the tree starting with the most similar pairs at the leaves and terminating at the root, using averaged atomic coordinates when merging more than 2 structures at an internal branch. At each merge, α -carbon equivalences are assigned using modified spatial and orientation probabilities (as in Rossmann and Argos (1976)). A matrix of probabilities for every possible α -carbon equivalence is computed, using probabilities averaged over all possible pairs of structures being merged. The best path through this matrix is assessed using a local dynamic programming step Smith and Waterman (1981) to select the most likely sequential C α equivalences. Again like some of the older methods, cycles of equivalence assignment followed by superposition are applied until the equivalence list is stable, and the process repeats for the next merge (May and Johnson, 1994; May and Johnson, 1995; May, 1996).

At the secondary structure level, Murthy (1984) describes a fast, two stage, superposition method, in which helices and strands are represented by their axial vectors. The first stage derives from Rossmann and Argos (1976) in which

rotational space is systematically sampled and at each step a matrix of angular orientation scores for the secondary structure vectors between each protein is produced. Each cell indexed by a pair of secondary elements from the two proteins is assigned a weighted score that is maximal for parallel vectors. Dynamic programming is then used to determine the best alignment and overall similarity score for each matrix, these being ranked and the highest selected as identifying the secondary structural equivalences. In the second stage, these equivalences are then used in another rotational search to achieve superposition by minimising the differences between vectors linking all pairs of elements in one protein and equivalent vectors in the second. The score is modified depending on how well vectors between equivalent secondary elements can be superposed. Finally, these modified scores are plotted as a function of the three Eulerian angles giving a contour map wherein strong structural similarities are revealed as peaks.

6.2 Using the relationships of internal features

Capturing the relationships between internal features is the most general and reliable approach to structure comparison but also computationally the most difficult. Its power derives from the use of relationships to capture the true 3D interaction of elements while still retaining useful intrinsic similarities in their encoded features — including the raw sequence data if desired.

Computational complications arise since, in general, the nature of relationships are not local in sequence and so violate the basic assumption of dynamic programming. Despite this, methods have been developed for this type of data that use dynamic programming. Two of the original attempts will be described below, one of which (**COMPARER**) uses a stochastic minimisation method to refine the matching of relationships while the other (**SSAP**) employs dynamic programming at two distinct levels.

6.2.1 The **COMPARER** program

Šali and Blundell (1990) recognized that the problem of comparing structures in terms of relationships was not directly amenable to conventional dynamic programming. Their program, **COMPARER**, compares proteins at various structural levels using a multiplicity of features and relationships. For each kind of structural element, features are compared and scored with weights into a matrix indexed by the two sequences. Relationship sets are analysed using simulated annealing to identify and weight elements participating in similar relationships in the two proteins, these scores being added into the matrix. Finally, the matrices for each structural level are summed, using weights to control the contribution of each structural level, and an overall alignment is generated using dynamic programming.

The authors also describe how they apply the method to multiple alignments. An initial series of pairwise structural alignments is used to construct a similarity tree (or the user can specify their own hierarchy). Multiple alignment proceeds by a sequence of pairwise alignments, in order of similarity following the topology of the tree, merging sub-alignments as necessary until all structures have been incorporated.

6.2.2 The SSAP program

The SSAP program (Taylor and Orengo, 1989b), and its derivatives (Taylor and Orengo, 1989a; Orengo and Taylor, 1990; Orengo and Taylor, 1993; Taylor *et al.*, 1994a), (see Orengo and Taylor (1996) for a review) uses a ‘double’ dynamic programming algorithm to manipulate two tiers of matrices⁵. A single upper matrix is used to score features directly and to accumulate alignment paths from the lower tier matrices, which are used to compare relationship sets of each possible pair of residues. The principal relationship employed uses a local structural environment about each residue comprising a simple reference frame defined by the geometry of the C_α atom. Every other residue is defined in this frame by a set of interatomic vectors. (Figure 11). Residue equivalences given by the resulting structural alignments are used directly to produce a weighted superposition by the algorithm of McLachlan (1979) using the alignment score at each equivalent position (Rippmann and Taylor, 1991).

Other relationships examined include interatomic distances, H-bond energies, virtual H-bonds extending through sheets, and disulphide bridges, while features include residue accessibility, secondary structure assignment, backbone angles, solvent accessible area, and sequence similarity (Taylor and Orengo, 1989a). Multiple features and relationships are scored using a weighted polynomial scoring function, with choice of features, relationships, and weights under user control.

The full double dynamic programming algorithm is computationally demanding, but Orengo and Taylor (1990) show that only a small subset of lower level comparisons is necessary — an aspect that has been exploited in later developments (Taylor, 1999b) (more fully described below). A local alignment version using a modified Smith and Waterman (1981) algorithm (Orengo and Taylor, 1993), and a multiple alignment version (Taylor *et al.*, 1994a) (using the progressive multiple sequence alignment algorithm of Taylor (1988)) were developed also. The latter generates a consensus structure by averaging vectors between equivalenced residues. Information gathered on structural variability of individual vectors and environments can be used, for example, to weight structurally conserved positions as more structures are added to the alignment.

SSAP can also align secondary structure elements (SSEs) (Orengo *et al.*, 1992) using secondary structure features (hydrophobicity, length, surface area) and rela-

⁵This older implementation of the algorithm will not be described in detail here since the current iterated algorithm will be described in the following section.

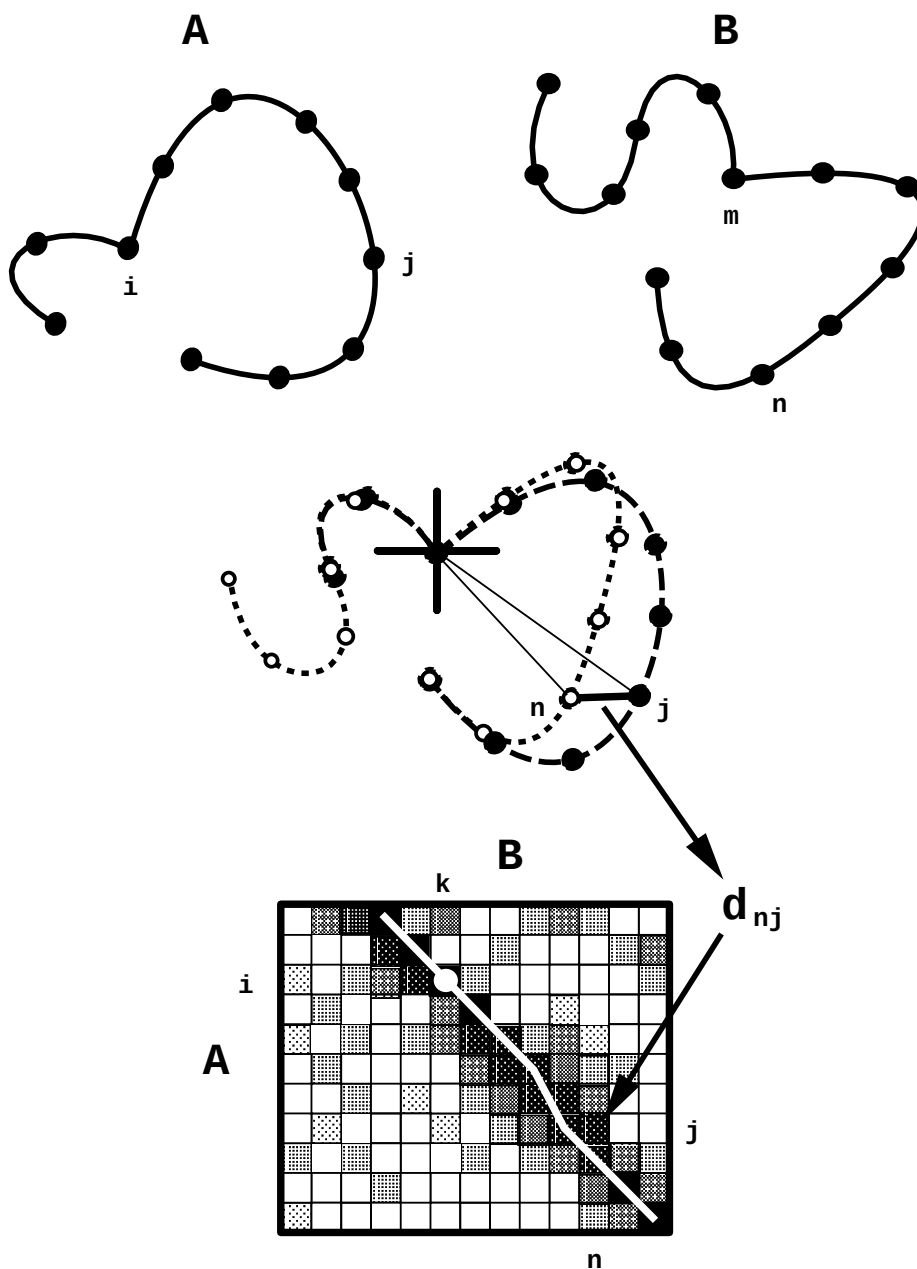


Figure 11: Two protein structures **A** and **B** are shown schematically. A pair of positions (i in **A** and m in **B**) is selected. Both structures are centered on i and m and orientated by a local measure (indicated by the large cross). In this superposition the relationship between all pairs of atoms (e.g. n and j) is quantified, either as a simple distance (d_{nj}) or by some more complex function. All pair values are stored in a matrix and an alignment (white trace) found. The arbitrary choice of equating i and m is circumvented by repeating the process for all possible i, m superpositions and pooling the results. In the **SSAP** algorithm a final alignment is extracted from the summed results by a second dynamic programming step.

tionships (buried area, overlap, tilt and interaxial angles, inter-element vectors). The resulting alignment of SSEs can then be used to constrain a subsequent residue level alignment. The method is fast and allows sorting of the structure databank into unique fold families (Orengo *et al.*, 1993).

6.3 Iterated Double Dynamic Programming

The program **SAP** (for Structure Alignment Program) described in this Section was derived from the related **SSAP** program (Taylor and Orengo, 1989b; Taylor and Orengo, 1989a) (Section 6.2.2) and is largely a simplification of its predecessor but is based on a refined iterative algorithm. The method is fully described here as some of its results are used below in Section 7 and Part III. The core comparison algorithm underlying both **SAP** (as well as **SSAP**, and also some sequence/structure comparison methods (Jones *et al.*, 1992; Taylor, 1997a)) is based on the dynamic programming algorithm.

6.3.1 Double Dynamic Programming

The computational difficulty in structure comparison programs like **SSAP** and **COMPARER** arises through trying to obtain a measure of similarity between two sets of internal relationships in different proteins. To compare the internal relationship of, say, residue i in protein **A** with a residue m in **B** relies on matching the individual relationships (such as $\{i, j\}$ in **A** with $\{m, n\}$ in **B**) (Figure 11). If this known (even for one such i, m pair) then the comparison problem would be solved before the first step was taken! To break this circularity, the following computational device was used: given the assumption that two residues (one from each of the two proteins) are equivalent, then how similar can their relationships (or structural environments) be made to appear while still retaining topological equivalence?

This aspect of the calculation is described in Figure 11 in which the score matrix is referred to as the *low-level* matrix (R). The scores along the best path through this matrix are then summed into a ‘master’ matrix (S), referred to as the *high-level* matrix. After all residue pairs have been considered and their path-scores summed in S , the best path is now found through S giving a best-of-the-best (or consensus) result. Representing the application of the dynamic programming algorithm as a matrix transform function $\mathcal{Z}$ that sets all matrix elements to zero except those that lie along the best path, then the full algorithm can be summarised as:

$$\mathcal{Z}(S) = \mathcal{Z}\left(\sum_i \sum_j \mathcal{Z}(R_{ij})\right) \quad (1)$$

where the sums are over all residues (i) in one protein with all residues (j) in the other.

The basic alignment (or Dynamic Programming) algorithm is thus applied at two distinct levels: a low-level to find the best score given that residue i is equivalent to j , and at a high-level to select which of all possible i, j pairs form the best alignment. This double level (combined with the basic algorithm) gave rise to the name “*Double Dynamic Programming*” (DDP).

6.3.2 Selection and Iteration

The DDP algorithm described above, requires a computation time proportional to the fourth power of the sequence length (for two proteins of equal length) as it performs an alignment for all residue pairs. To circumvent this severe requirement, some simple heuristics were devised based on the principle that comparing the environment of all residue pairs is not necessary. By considering local structure and environment, many residue (indeed most) pairs can be neglected. This selection is based on secondary structure state (one would not normally want to compare an α -helix with a β -strand) and burial (those with a similar secondary structure and degree of burial are selected) but a component based on the amino acid identity can also be used, giving any sequence similarity a chance to contribute.

An iterated algorithm was implemented previously (Orengo and Taylor, 1990), using heuristics on the first cycle to make a selection of a large number of potentially similar residue pairs. In the reformulated algorithm, a small selection (typically 20–30) pairs are selected initially and gradually increased over several iterations. This initial sparse sampling can, however, be unrepresentative of the truly equivalent pairs and to avoid this problem, continuity through the early sparse cycles was maintained by using the initial rough similarity score matrix (referred to as the *bias* matrix) as a base for incremental revision. (Figure 12). As the cycles progress, the selection of pairs becomes increasingly determined by the dominant alignment, approaching (or attaining) by the final cycle, a self-consistent state in which the alignment has been calculated predominantly (or completely) from pairs of residues that lie on the alignment.

6.3.3 Sampling alternate alignments

A useful ‘spin-off’ from the iterated DDP approach is to augment, or bias, the evolving selection of pairings (referred to as the current selection). This can be done using external information such as sequence or structural patterns (Jonassen *et al.*, 1999), or by adding random displacements to the scores on which the selections are based. This latter approach introduces some of the aspects of the stochastic methods discussed above (Šali and Blundell, 1990) (Section 6.2.1) and is equivalent to sampling the population of high scoring alternate alignments.

Knowledge of the distribution of sub-alignments gives an idea of how unique the highest scoring alignment is (and indeed, whether this best alignment was

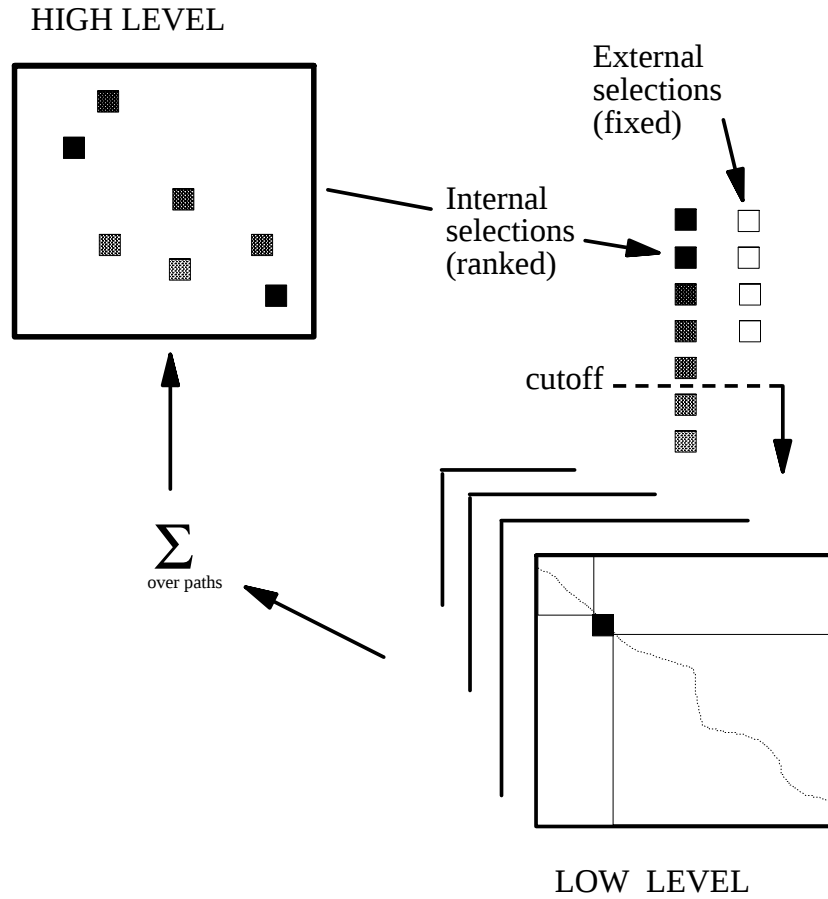


Figure 12: **Outline of the iterated double dynamic programming algorithm.** Values from the *HIGH LEVEL* score matrix are ranked and a pre-specified number (represented by the dashed *cutoff* line) are passed to the *LOW LEVEL* for evaluation. These are joined by a fixed number of externally specified pair-selections. The resulting alignment paths are summed back into the *HIGH LEVEL* score matrix and, after normalisation of the values, a new selection is made. Five cycles of iteration are typical.

found by the ‘one-shot’ algorithm). If the best alignment is unique (few similar scoring alternatives) then it can be treated with confidence whereas if there are a lot of equally scoring alternatives then care must be taken in interpreting it in detail. Ways in which this can help also in assessing the significance of an alignment score are discussed in Section 7.

7 Assessment of Significance

Like sequence alignment methods, almost all of the methods discussed above will produce a match when presented with two structures — whether these structures share any similarity or not. An important aspect of structure comparison is to decide when a match is significant. This is difficult as we have seen that, beyond close similarity, there is no uniquely correct structural alignment of two proteins and different alignments are achieved depending on which biological properties and relations are emphasised in the comparison (Taylor and Orengo, 1989b; Godzik, 1996; May and Johnson, 1994; May, 1996).

For a proper statistical assessment, The scoring found for a structure comparison must be compared against what is expected by chance. This is often implemented as what is expected by aligning random structures, or using fragments of non-related proteins.

7.1 Score distributions from known structures

Alexandrov and Go (1994) made an analysis for finding the significance of similar pairs of proteins using their program SARF. For a fixed length L , they picked up all fragment pairs of this length in two unrelated structures, and found the value R_L such that only 1% of pairs have smaller RMSD. Similarly, Russell (1998), made an analysis using distance RMSD, related to his method for detecting side-chain patterns. Random pairs of structures with different folds were chosen, and random patterns of two to six patterns were derived.

Alexandrov and Fischer (1996) and Holm and Sander (1993b) used a Z-value statistic to measure significance whereas Gibrat et al. (Gibrat *et al.*, 1996), in their VAST program, compute a P-value for an alignment based on how many secondary structure elements are aligned as compared with the chance of aligning elements randomly. Levitt and Gerstein (1998), made a comparison of the scoring of their iterated dynamic programming/superposition program to RMSD. The P-value of a score S for fixed N (number of matched residues) can be found by fitting to an extreme-value distribution. The same statistics can be developed for use of RMSD and both can then be compared by a method of Brenner *et al.* (1998) in which the E-values of each structure pair giving 1% false-positives was taken.

7.2 Random structural models

In sequence comparison, the generation of a set of random models is easily achieved by generating random sequences either as a Markov process or through shuffling the native sequences. However, no such simple method can be used for structures and the best random model against which comparison scores should be compared depends on the degree to which the inherent nonrandom features of protein structure in general should be considered significant (Taylor, 1997b). Random chains can be generated and compared (McLachlan, 1984) but the best random models would be those generated with secondary structures. Ideally, these models should be calculated for each comparison to match the length of the native comparison and the secondary structure composition (Aszódi and Taylor, 1994b). However, these models are complex to generate and cannot be ‘tailor-made’ for each individual comparison without excessive computation.

Models involving symmetry operations on the protein (reversal and reflection) can be used in situations where the comparison method restricts its calculation to the α -carbon atoms of the protein since the arrangement of the other main-chain atoms is directional (Taylor, 1997b; Maiorov and Crippen, 1994). Considering just α -carbons, the conformations of local structural features (such as secondary structure and their chirality of connection) in the reversed chain is virtually indistinguishable from a forward running ‘native’ chain. This principle of reversal applies equally at the level of the sequence and has been used previously to provide a random model for sequence pattern matching (Taylor, 1986b; Taylor, 1998). In both sequence and structural data the reversed model preserves the length and composition of the protein, including directionally symmetric correlations associated with secondary structure, while additionally in the reversed structural model, the bulk properties of packing density and inertial axes are also preserved. The latter are difficult to maintain in randomly generated structures (Aszódi and Taylor, 1994a).

The reflected chain is clearly not an ideal model for proteins as they contain both large and small scale chiral features which will change hand under reflection. However, the use of greatly simplified lattice models avoids this problem and based on this analysis, Maiorov and Crippen (1994) proposed a definition of the significance of RMSD in which they take two conformers to be intrinsically similar if their RMSD is smaller than that when one of them is mirror inverted.

7.3 Randomised alignment models

In general, the closer the random model is to preserving the properties of the native proteins, the more difficult it becomes to generate plausible alternatives. This problem is particularly acute for the reversed-chain random model discussed above since, for any given protein, there is only one. This problem can be partially circumvented, however, at the stage of calculating the alignment. At this point

the alignment with each random model can be expanded into a population of variants by introducing ‘noise’ into the score matrix and repeating the calculation of the alignment path from each noisy matrix. This generates a family of near-optimal subalignments and the spread of scores for this population can provide a measure of the stability or uniqueness of the answer. An advantage of this approach is that it can be applied not only to structures belonging to the set of randomised models but also to the native structure itself and the two resulting score distributions can be tested statistically to see if they are distinct.

If there is sufficient ‘noise’ introduced into the alignment method, and the population is large enough, then almost all reasonable alignments for a pair of proteins can be sampled. Plotting these solutions by their number of aligned positions against RMSD revealed a ‘cloud’ of points which was diffuse at high RMSD but had a sharp boundary on its lower edge. This edge represents the limit, for a given number of aligned positions, below which a smaller RMSD cannot be found. As judged by the sharp edge to the distribution, this limit is not restricted by the method of comparing the proteins and so provides an absolute standard against which other methods can be compared. For a few protein pairs, the results of other methods (gathered by Godzik (1996)) were plotted and compared to these lines. Most of these results were found to lie above the line, indicating that the optimal solution in terms of minimum RMSD had not been attained. Only a few results lay on the line and these mostly involved a smaller number of equivalent positions. It should be noted that assessing methods by the use of the RMSD value is sometimes unfair since for many of the methods, their aim is not to minimise the RMSD value.

7.4 Scoring and biological significance

When a structure is compared to every other structure (or to a representative selection), then scores will result ranging from the clear relationships of homologous proteins to a large number of poor scores for obviously unrelated pairs. Between these extremes lies a “twilight” zone within which it is very difficult to assess the significance of the score. This problem is exacerbated because many proteins contain similar substructures, such as secondary and super-secondary structures and the problem is to decide when a similarity is just a consequence of being protein-like and when it indicates a more specific relationship between the two proteins.

Because of its common currency, most considerations of this problem have focused on the significance of the RMSD measure based on comparison of proteins or protein fragments of equal length (see above). Others, such as the DALI method, have adopted a similar approach based on the scores achieved over matches of protein fragments (Holm and Sander, 1993b). Both these approaches require that the selected fragments are unrelated to the proteins being assessed, however, this raises the problem of what criterion can be used to make

this distinction and, in principle, it should not be a weaker method than that used for the current comparison. It is not acceptable, either, to consider completely unrelated proteins since, to take an extreme example, if the two proteins being compared contained only α -helices and the clearly unrelated control set contained only β -structure, then the two α proteins would appear more related than they should do.

An alternate approach to this problem is to use the reversed structure (as described above). When this is matched against the structure databank a similar range of scores should result — since the reversed structure has exactly the same length, overall shape, and secondary structure content as the native probe. What will be lost is any specific overall similarity to proteins that are homologous to the native probe. In addition, if the probe structure is a particularly simple fold (such as four α -helices) then the reversed structure will also embody this property so a specific match will need to capture more than a few matched helices to gain significantly over the background of scores derived from the reversed structure.

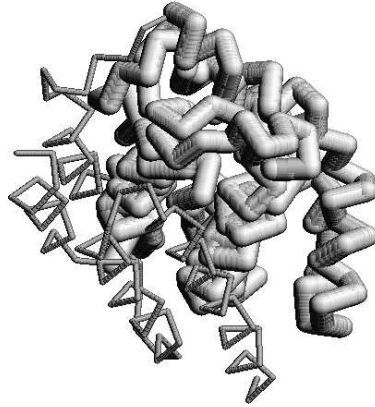
7.5 Examples

7.5.1 Distant globin similarities

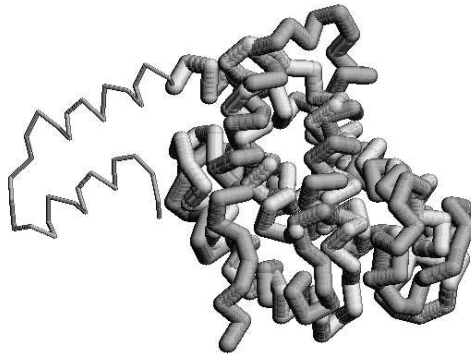
A globin-like fold is also found in the plant phycocyanin proteins which have the same core fold of six helices with two ‘extra’ ones preceeding this core (Pastore and Lesk, 1990). These proteins have no significant sequence similarity and only a vague relationship: both bind co-factors, however, the phycocyanins are electron transport proteins specific to the photosynthetic complex and do not bind oxygen as do all the globins.

The globin fold has also been found in the bacterial toxin protein colicin-A. As with the phycocyanins, this is larger than the globins, but in this relationship, the equivalent fold must be extracted from an otherwise well packed bundle of 8 helices (Holm and Sander, 1993a; Orengo and Taylor, 1993). Here no amount of imagination can lead to a plausible functional or evolutionary link with the globins (or the phycocyanins).

Progressing in the other direction, many small folds can exhibit similarity with part of the globin fold — in the extreme, this might involve matching a single α -helix. A relationship of a small protein with the globins that has been considered ‘significant’ was noted in the DNA-binding domain of the bacterial repressor proteins Subbiah *et al.* (1993). This consists essentially of only three helices but, overall, these adopt the same fold as part of the globins. The authors suggested a possible evolutionary relationship here since the DNA- and hæm-binding sites are located in similar parts of the structure — but this is rather speculative.



(a) colicin



(b) phycocyanin

Figure 13: **The globin fold in colicin-A and phycocyanin.** The two structures are drawn to show their backbone as linked α -carbons with the region corresponding to the globin fold drawn more thickly. (a) Colicin [1colA], which has extra helices towards the carboxy terminus. The core region matched 97 residues with an RMS deviation of 3.2Å. (a) Phycocyanin [1cpcA], which has two extra helices on the amino terminus. Both structures were compared against the hemoglobin structure 1h1b (sea cucumber). The core region matched 85 residues with an RMS deviation of 5.4Å.

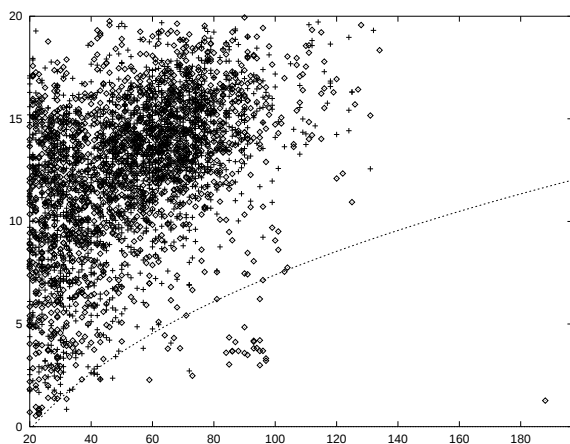
7.5.2 Assessment against chain reversal model

The three globin fold similarities were reexamined using the reversed chain as a background score model. A globin fold was scanned against a non-redundant selection of the Protein Structure Databank (PDB)⁶, and found all globins and all phycocyanins as significant matches. Similarly, a phycocyanin probe found all phycocyanins and globins.

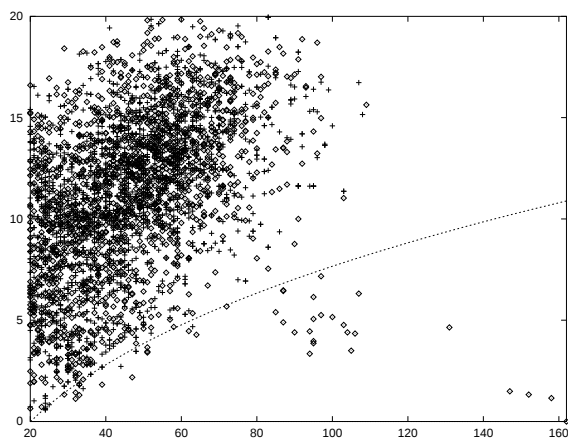
The globin/colicin-A relationship was tested using the reversed colicin structure as a control and was found to retain its significance (Figure 14(a)) with the globin finding colicin and colicin finding the globins. However, while colicin could find a phycocyanin, a phycocyanin probe did not find colicin at a significant level. (Figure 14(b)).

When tested in a similar way, the relationship between the bacteriophage repressor and the globins was not found to be significant when using a globin as a probe, however, with the repressor as a probe a number of globins were found to lie just on the border-line of significance.

⁶This selection was made by choosing one representative for all sequences that share greater than 50% sequence identity. The member taken to represent each family fulfilled a variety of criteria but generally had the best resolution and the lowest average B-value (an indicator of refinement quality). Details can be found in Taylor (1997a).



(a) colicin



(b) phycocyanin

Figure 14: **Structural searches with colicin and phycocyanin.** RMS deviation is plotted against number of residues aligned (diamonds). The structures match themselves (point on the lower right corner) and any homologues (clustered in the lower right corner). The probe structure was then reversed and rescanned (crosses). from these results a line (dashed) was fitted that excludes 99% of the reversed matches. (a) Colicin [1colA], has only one homologue and the cluster of matches around 2–4Å RMS and 80–100 residues include globins and phycocyanins. (b) Phycocyanin [1cpcA], has a few homologues and matches the globins in a cluster around 4–6Å and 80–110 residues. Colicin was not found at a significant level.

8 Protein Structure Classification

8.1 Introduction

We are currently faced with a flood of protein 3D structure data: as we enter this Millennium, there are 11515 entries⁷ in the Protein Data Bank (PDB) (<http://www.rcsb.org/pdb>). Furthermore, the number of structures in this database is doubling almost every 18 months.

Despite this flood of data, increasingly efficient and robust methods for protein 3D structure comparison have made it feasible to perform all-against-all comparisons of all known 3D structures. (for a review, see Holm and Sander (1994b) and Orengo (1994) and also the previous Sections for details of the methods). These exhaustive comparisons reveal that proteins can share a common fold despite lacking any ‘significant’ sequence identity (Section 7.5) and, furthermore, proteins with the same fold may have different functions. Their main aim, however, is to try and bring some order into the description of protein structure by imposing a classification.

In this Section we investigate some of these approaches and ask whether the evolutionary model, that is used when there is clear sequence similarity, can be extrapolated into these more tentative relationships and whether attempts at classification lead to a greater understanding of protein structure.

8.1.1 Practical applications

The following list gives some areas in which (even a rough) classification has proved useful, and some areas in which it should still be of use.

1. Classification helps us to understand protein evolution since structure is better conserved than sequence (Lesk and Chothia, 1980; Chothia and Lesk, 1986; Chothia and Lesk, 1987).
2. It is useful to describe protein fold-space and maybe answer the question: how many folds are there in Nature?
3. With the increasing number of proteins for which an experimentally determined 3D structure is available it is helpful to have an ordered collection of all known folds to ascertain whether a new structure is in fact a novel fold.
4. Classification aids our understanding of the relationships between 3D structure and function such as in enzymes (Thornton *et al.*, 1999).

⁷This value overestimates the number of different structures as there is a high level of redundancy in the PDB with the same or slightly modified protein being seen in different structures or under different conditions. There is also obvious bias in structure determination towards small, single domain proteins amenable to such experimental investigation and also towards those proteins deemed to be of interest (e.g. enzymes or DNA-binding proteins).

5. If we have a fixed number of different structure types, then predicting a protein structure from sequence will involve a finite search. (Section 12).
6. Classification makes protein 3D structure data more accessible to biologists and other non-specialists.

8.1.2 Genome applications

In 1995 the complete genome sequence (i.e. the entire DNA) of a bacterium became available. This was the first organism larger than a virus to have its genome determined. Since then almost another 50 microbes have had their genomes sequenced and, recently, we have seen the much larger genomes of three ‘higher’ organisms (brewer’s yeast, a nematode worm and a fruit fly) and the human genome is expected this year (2000).

A major challenge in the post-genome-sequence era lies in the functional annotation of genomes: assignment of function to each gene product. In the absence of an experimentally defined function, the most reliable method for predicting function is on the basis of sequence similarity to proteins of known function. If a protein of unknown function can be assigned to a protein fold then this can add even more value in terms of structure-function relationships. It is for this reason also that 3D structures are now being determined experimentally for proteins with no known function. Clearly, classification of protein folds is key to functional annotation of genomes.

A related issue is that of target selection for structural genomics. The aim of structural genomics is to assign a 3D structure to all the proteins encoded by a genome. However, whatever the eventual number of genes in the human genome, it is clear that is not feasible to determine experimentally the 3D structure of every human protein. Instead, it should be possible to identify those proteins whose structures will reveal new folds. By definition, given a complete set of folds, it will then be possible to model all proteins for which an experimentally defined 3D structure is not available.

8.2 Practical approaches to classification

One dictionary defines the term ‘classification’ as the “systematic placement in categories” (Collins Paperback English Dictionary (1993)). But to be systematic, the placement must be made according to established criteria for similarity that describes the extent of resemblance between the objects. Clearly, the question of how to define similarity is not trivial and can be highly subjective. Indeed, all the various ways in which proteins can be compared that were described in the previous Sections, will give rise to systematic but differing classifications.

We will discuss the three most popular classifications, all of which are accessible via the World Wide Web (WWW)⁸. In summary, these are:

1. SCOP: a Structural Classification of Proteins database (Murzin *et al.*, 1995; Hubbard *et al.*, 1997) which is essentially a manual classification.
2. CATH (Orengo *et al.*, 1997) which is constructed using both manual and automated approaches.
3. FSSP (Holm and Sander, 1997; Holm and Sander, 1998) which is built in a totally automated fashion.

All three classifications use a hierarchical data structure with a nested set of partitions grouping similar proteins.

8.2.1 Automated approaches to classification

Given an approach to define, preferably, a topological equivalence between a pair of 3D structures we need a measure to describe their extent of similarity or distance (a metric). Most metrics specify the pairwise DISSimilarity: for example, the most common dissimilarity measure is the root-mean-square deviation (RMSD) after rigid-body superposition⁹.

Unfortunately, as we have seen in the previous Sections, unlike amino acid sequence alignment, the problem of 3D structure alignment is not trivial. Although sequence alignment using dynamic programming guarantees the optimal solution (mathematically but perhaps not biologically), the comparison of 3D coordinate data is not as well defined as the comparison of 1D strings of amino acids. This gives even more scope for the measures produced automatically to differ, as an alignment between 3D structures depends on the nature of the objective function. For instance, intermolecular distances might be minimised in a rigid-body superposition (*e.g.* (May and Johnson, 1994; May and Johnson, 1995)), or they might be compared in a pairwise manner, as in the **SSAP** (Taylor and Orengo, 1989b) and **Dali** (Holm and Sander, 1993b) programs.

Another consideration for structure alignment, say by superposition, is the balance between the number of topological equivalences and the attendant RMSD (May, 1996) in this case, the goal is to maximise the number of equivalences while simultaneously minimise the associated RMSD. The question arises then as to how to identify the alignment with the most meaningful compromise between the two factors (May, 1996; Taylor, 1999b).

⁸Structure database web sites:

SCOP = <http://scop.mrc-lmb.cam.ac.uk/scop/>,

CATH = <http://www.biochem.ucl.ac.uk/bsm/cath/>,

FSSP = <http://www2.ebi.ac.uk/dali/fssp/>. The latter is generated by the program **Dali**.

⁹It is important to specify over which atoms the RMSD is calculated. In the current discussion it can be assumed that only the main chain α -carbon atoms are considered but any different choice obviously affects the result.

8.3 Organisation of the classifications

8.3.1 The unit of classification

Despite the differing philosophies behind the three classifications, (SCOP, CATH and FSSP) there is consensus on the unit of classification: the protein domain. (Section 2.7). There are several algorithms for domain identification from coordinates (Taylor, 1999c; Holm and Sander, 1994a; Swindells, 1995) but even a structure-based definition is often non-trivial. For instance, there are often extensive interfaces between domains leading to ambiguity about the appropriate level of granularity for domain definition. Another complication lies in the fact that domains can comprise sequential (continuous domains) and non-sequential (discontinuous domains) parts of the polypeptide chain (Figure 7). Continuous domains are easier to identify than discontinuous ones (Jones et al. (1998)). Not surprisingly, differences in domain assignment have been shown to be an important factor between the classification schemes (Hadley and Jones, 1995) although other, less-structural, criteria are involved such as folding (independently folding units) or function (functional units).

8.3.2 Hierarchical organisation

Although all three major classifications agree on a hierarchical paradigm, they differ in the detailed organisation. For example, the top level of the hierarchy in SCOP and CATH is protein class. However, SCOP and CATH differ in the number of classes used. While SCOP uses the original four classes of Levitt and Chothia (1976), CATH merges the α/β and $\alpha+\beta$ classes into a single one.

CATH has a unique level within its hierarchy: architecture. Architecture is the overall shape of a domain as defined by the packing of the secondary structure elements but ignoring their connectivity. The current release of CATH (version 1.6 June 1999) consists of 35 architectures which have been assigned by eye. (A more systematic approach will be outlined in Section 10).

All three classifications agree on a fold level. The fold of a protein describes its architecture together with its topological connections. However, there is a difference in how folds are assigned. For instance, it is done automatically in CATH on the basis of structure similarity score derived by SSAP (Taylor and Orengo, 1989b). However, in SCOP, fold definition is done by eye.

Although proteins are grouped into families and superfamilies, once again the operational definition of these terms can vary. Families comprise proteins believed to be homologous i.e. those related by divergent evolution from a common ancestor. Clear evolutionary relationship is usually assigned on the basis of significant sequence identity. Here there are differences: SCOP uses a threshold of $\geq 30\%$ sequence identity while CATH uses $\geq 35\%$. Of course, in those cases where family membership is assigned on the basis of common fold and function, in the absence of significant sequence identity (*e.g.* as with the globin examples

discussed in Section 7.5), then a problem remains in definition of a common fold. Superfamilies comprise proteins deemed to share a probable common evolutionary origin on the basis of a common fold and often function but in the absence of significant sequence identity. (A detailed comparison of SCOP, CATH and FSSP is described in Hadley and Jones (1995).)

8.3.3 Hierarchical classification

Hierarchical organisation is a key concept not only in protein structure and its classification but also across all biology. For instance, the formal system for classifying and naming organisms — Linnaean taxonomy — is based on a simple hierarchical structure. Furthermore, hierarchical classification is the most frequently used method of cluster analysis. The result of a hierarchical classification of a set of objects is a tree resembling a phylogenetic (evolutionary) tree. A tree used for phylogeny inference is almost always tested in terms of the support for a tree representation and individual clusters contained within. One of the most popular methods to attach confidence limits on phylogenies is via bootstrap replicates (Felsenstein (1985)).

Surprisingly, until recently, trees derived from protein 3D structures had not been assessed in such a way. Recently, May (1999a,b) used a jackknife test to identify meaningful 3D structure-based trees — defining a meaningful tree as one where all the clusters are found to be reliable according to the jackknife test. For example, applying this test to the relationships between small $\beta\alpha$ proteins (Taylor *et al.*, 1994a) found that 3 of the 9 clusters contained within the tree were unreliable according to the jackknife test (May (1999a)). Such an approach allows the investigation of the suitability of a structure (dis)similarity measure for hierarchical classification of protein folds (May (1999a)).

8.4 Remaining Problems

8.4.1 What questions does classification help us to answer?

The current version of CATH (version 1.6 June 1999) contains 672 folds while there are 520 in SCOP (release 1.48 Nov 1999) and with these large well organised classification schemes, it is possible to compile population statistics of 3D structures. Such analysis has shown that some folds occur more often than others such as the TIM barrel (Figure 5). This structure was first seen in triose phosphate isomerase (TIM) (Banner *et al.*, 1975), an enzyme in the key metabolic pathway glycolysis. In fact, not only are all the glycolytic enzymes α/β structures but also the last enzyme of the pathway, pyruvate kinase, contains another TIM barrel domain. Approximately 10% of all known enzyme 3D structures have a TIM barrel fold despite having different amino acid sequences and different functions (for a recent review, see Reardon and Farber (1995)). Along with a few other

folds, the TIM barrel has been termed a “superfold” (Orengo *et al.*, 1994): a fold common to at least three non-homologous proteins (i.e. with no significant sequence identity).

Classification has made it possible to explore global relationships between protein 3D structure and function. For example, originally Nishikawa and Ooi (1974a), and more recently, Martin *et al.* (1998) have shown that most enzymes have α/β folds. It is also possible to identify densely populated regions of fold space — referred to as ‘attractors’ in Holm and Sander (1996)).

Brenner *et al.* (1998) have used the SCOP classification as a “standard of truth” to evaluate the effectiveness of sequence alignment methods. The idea is that protein relationships defined according to 3D structure and function can serve to benchmark methods that match proteins on the basis of only sequence similarity.

8.4.2 Questions raised by classification

Analysis of the various classifications has helped us to refine our ideas of protein 3D structure similarity. However, further questions are also raised:

- How might we best represent similarity relationships?
- Is a hierarchy the best model?
- Is it possible to reach consensus on terminology such as how to define a architecture, fold, and family?
- Most importantly, might classification be made less subjective?

SCOP defines a separate class for multi-domain α and β class proteins and for folds consisting of more than one domain of different classes. Is it feasible to improve the classification of such folds? As we have seen, multidomain proteins are subdivided into domains for the purpose of all current classifications. By definition, however, the function of multidomain proteins is a property of the entire structure. This problem is only going to get worse as the continuing advances in technologies for protein 3D structure determination mean that more and more structures will become available for large, multidomain proteins. Similarly, the focus of structure determination is moving towards protein-protein complexes such as those involved in transcription or signal transduction.

Not surprisingly, there has been much speculation as to the total number of protein folds in nature. One, often quoted, estimate is that there are 1000 folds (Chothia, 1992). However, a recent calculation puts the figure at around 2000 (Govindarajan *et al.* (1999)). In fact, the only area of agreement within the community is that the number of protein folds in nature is finite! Whatever the actual answer, we need to consider the question of how many folds remain to be seen. Of course, this is not just an academic question given the investment

required for structural genomics. Clearly, classification helps to define sparsely populated regions of fold space and so can help to direct protein 3D structure determination.

Organising known protein 3D structures into classifications has only served to emphasize the paucity of membrane protein 3D structure data. For instance, excluding proteins only anchored in the membrane, there are only 10 known membrane protein folds according to SCOP (release 1.48 Nov 1999). Recently, Jones and Taylor (1999) have suggested the existence of transmembrane protein superfolds.

Classification of proteins on the basis of common fold and function informs hypotheses about how proteins evolve new functions. What is the relationship between protein fold and folding pathway? In a series of papers, Efimov (for example, see Efimov (1997)) has classified protein folds by constructing what he describes as "structural trees". The root of each tree is a motif common between all folds described therein. Each fold can then be described in terms of stepwise addition of secondary structure elements, on the basis of the rules of protein 3D structure, to the basic motif. Efimov has suggested that not only might the core motifs represent nuclei in protein folding but also that the pathways of their stepwise elaboration could correspond to folding mechanisms.

8.4.3 Future prospects

Since we do not yet have a complete library of protein folds, any classification can only be a snapshot of a dynamic situation and this means that the classifications need constant updating. This emphasises an important difference between the three classifications: FSSP, because its construction is entirely automated and so is always up to date; however, SCOP and CATH need considerable human input and so are behind the latest release of the structure data. More fundamentally, as we have seen, there is an unacceptable level of disagreement about the usage of certain terms and what is important in a classification. It is to be expected then that even when a complete set of protein folds is available there will be many discrepancies between classifications.

In Rutherford's division of science, protein fold classification currently bears a greater similarity to stamp-collecting than to physics! In many ways, it represents little more than fact accumulation and sorting. Indeed, one might wonder whether attempts to classify protein folds are simply a reflection of an innate human desire to impose order and certainty on an otherwise unconnected collection of folds? What we lack at the moment is a general physical theory to synthesise the current data. This might come from a better understanding of how a 1D amino acid sequence specifies a particular 3D structure (the "protein folding problem") but at the moment we can do little more than catalogue each new protein 3D structure and hope, as occurred in Natural History of the mid-nineteenth century, for the arrival of a new Darwin to guide us out of the wilderness!

Part III

Geometric Abstractions and Topology

*For the want of a bond, a strand was missed,
For the want of a strand, a sheet was missed,
For the want of a sheet, a fold was missed,
All for the want of a hydrogen-bond.*

Adaped from the nursery rhyme *the Horseshoe Nail*
(with apologies to Anon.)

9 Simplified Geometries

9.1 Structure Representations

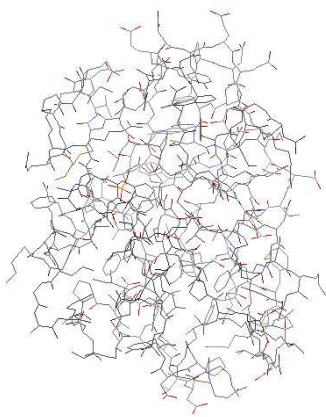
9.1.1 From bonds to cartoons

Through the previous descriptions of structure and comparison, proteins have been represented in a variety of ways using different levels of detail. Although little of it has been seen hitherto, the full representation of proteins has all atom coordinates specified including hydrogens. For most X-ray analyses of structure, however, the hydrogen positions are not normally visible and the standard representation is generally to use all heavy (non-H) atoms (Figure 15(a)). While this level of representation is required for detailed analysis of substrate binding, packing and catalysis, in the present work, we have concentrated more on the overall fold of the protein and for this a representation of the protein backbone path is usually sufficient. This can be shown in many ways: some of which incorporate features derived from the more detailed levels: such as secondary structure.

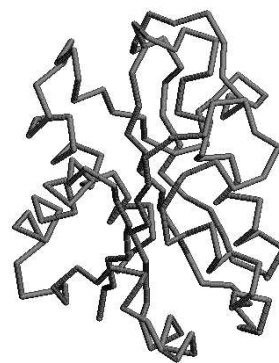
The simplest representation is to connect a central atom in each residue (and for this the α -carbon is the obvious choice) resulting in a trace that shows the overall fold of the protein clearly and in which secondary structure (if present) can also be seen (Figure 15(b)). This trace can be smoothed to different degrees to simplify ‘unimportant’ details in surface loops or the secondary structures can be emphasised by using a more symbolic representation (Richardson, 1985). This can be done without explicit definition of the secondary structures — using the orientation of the (flat) peptide plane ($> N - C <$) to guide the surface of a ribbon representation (Figure 15(c)) or with explicit secondary structures resulting in a similar representation but now the β -strand components have been ‘labeled’ with an arrowhead (Sklenar *et al.*, 1989; Carson, 1991; Thomas, 1994). The definitions of secondary structure used in the latter representation should also have been generated by an ‘expert’ (usually the scientist who determined the structure) or by an automatic algorithm that has explicitly considered H-bonding networks (such as the DSSP program (Kabsch and Sander, 1983)). These representations are shown together in Figure 15 for comparison. Each was generated from the program RASMOL (Sayle and Milner-White, 1995) which is principally intended to display these representations interactively. Finer quality but static representations can be generated from other programs such as Molscript (Kraulis, 1991), some examples of which can be seen in Figure 3 and Figure 5 in Section 2.

9.1.2 From 3-D to 2-D

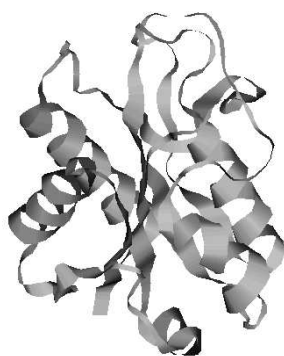
Secondary structures are extended (helical) objects and, because of their linear axis, often pack in a roughly aligned manner as in a bundle of rods. To a first approximation this allows the structure of proteins to be displayed in a very



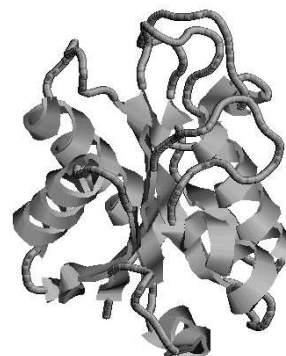
(a) All-atom model



(b) α -carbon trace



(c) Ribbon trace



(d) Cartoon representation

Figure 15: **Protein structures representations** A small β/α protein (flavodoxin) is shown in four representations. (a) showing bonds between all non-hydrogen atoms, (b) with lines connecting sequential α -carbon atoms, (c) as a flat trace (ribbon), drawn to follow the orientation of the peptide planes, (d) with explicit secondary structure definitions represented by 'cartoon' objects. The figures were produced by the program RASMOL.

simplified manner by neglecting the extended dimension and portraying only the ends of the ‘rods’. In this representation protein structures appear as layers of packed secondary structure (Figure 4). Typically, β on β (the β -sandwich class) or a β -layer between two α -layers (the alternating β/α class). The layered structure is clear in the preceding classes because of the regularity imposed by the hydrogen-bonded β -sheets. However, this constraint is not present in the all- α class which adopt a less regular variety of forms.

This form of representation is typically referred to as a ‘topology cartoon’ and has been used extensively to describe protein folds since some of the earliest analyses of structure (Sternberg and Thornton, 1977b; Sternberg and Thornton, 1977a; Nagano, 1977). It has also formed the basis for semi-automatic (Flores *et al.*, 1994) and fully-automatic analyses of proteins at the ‘topological’ level (Sternberg *et al.*, 1985; Rawlings *et al.*, 1985; Rawlings *et al.*, 1986; Clark *et al.*, 1991; Gilbert *et al.*, in Press; Gilbert *et al.*, 1999) including the basis of a text string description of structure (Flower, 1998).

10 Stick Representation

Representation of secondary structures as line segments introduces a great saving in the structural description of proteins without a significant loss of detail. The information that is discarded is the phase of the helix or strand relative to the rest of the protein along with whatever degree of detail has been discarded from the loop regions connecting the secondary structures. As illustrated above (Figure 15), the latter can range from none (in which the α -carbon trace is maintained through the loops) through varying degrees of smoothing to the situation in which the link between secondary structures is represented only by an abstract line or curve.

This economy of description has resulted in great savings in computational time in many of the various structure comparison methods described in Part II. In general the number of points is reduced by ten-fold and for algorithms that typically require execution times with, at best, cubic or quadratic order dependency on the number of points, then savings can be considerable. Consequently, it is at this level of representation — at which greatest simplification has been achieved with least loss of structural information — that it is convenient to gain an overview of the full range of protein structure and to devise ways in which it can be systematically represented and compared.

10.1 Secondary structure line-segments

10.1.1 Problems with current criteria

One of the problems that bedevils the analysis of protein structure at the level of secondary structures is to find a robust definition of secondary structure. As

the opening rhyme to this part emphasises, trivial differences at the atomic level can propagate upwards to become obvious differences at the higher level of representation. Taking this verse literally, a difference of as little as a fraction of an Ångström in the position of a main-chain hydrogen-bonding group might lead to the failure of an algorithm to recognise a potential hydrogen-bond. This might then leave a β -strand (on the edge of the sheet) to be too short to be incorporated into the sheet which could lead to a secondary structure representation with one less element between otherwise identical proteins.

One of the (few) advantages of a manual definition of secondary structure is that experts ‘gloss-over’ these minor aberrations and tend to make a more regular or ‘tidy’ definition of secondary structure. While good for an overview, if one is analysing disruptions in secondary structure then this is not a very useful approach. To minimise these difficulties, automatic methods tend to have a flexible definition of hydrogen bonding and also tend to base their definition on larger scale structures — such as hydrogen-bonded ladders (as in as the DSSP program of Kabsch and Sander (1983)) giving some degree of robustness.

A further problem, not well dealt with either by ‘eye’ (or automatically) is in deciding what the secondary structure is when there are only a few hydrogen-bonds involved. This might seem to be simple since the hydrogen-bonds are discrete: progressing through the various helices of 3_{10} , α and π in steps of one residue in the nearest bonded neighbour. However what can be made of the following pathological example in which each of the three helix types follows in progression (Figure 16). Clearly one could define three different helix types but the problem is that each overlaps each other in extent — in other words: although each helix only has one bond, this bond bridges a number of residues. The problem is further compounded by the ability of hydrogen-bonds to bifurcate and have two bonding partners!

The following section describes a simple physical method to avoid some of these problems while retaining a working definition close to what would be defined by an expert. It will be described in reasonable detail as it will be used later in this section for the further automatic analysis of protein structure in terms of secondary structure line segments.

10.1.2 Line segments from inertial axes

As discussed above, analysis of proteins in terms of the geometry of their secondary structure line segments depends on having robust definitions of secondary structure, which despite automatic approaches, are often sensitive to structure quality. For the methods described below, this area of ambiguity can be largely avoided by relying on a purely geometric definition of line segments. The axis of a secondary structure is typically taken as the line with minimum deviation (least-squares) from the α -carbons and this can be found as the principle axis of the equivalent inertial ellipsoid (Taylor *et al.*, 1983). More generally, if the size

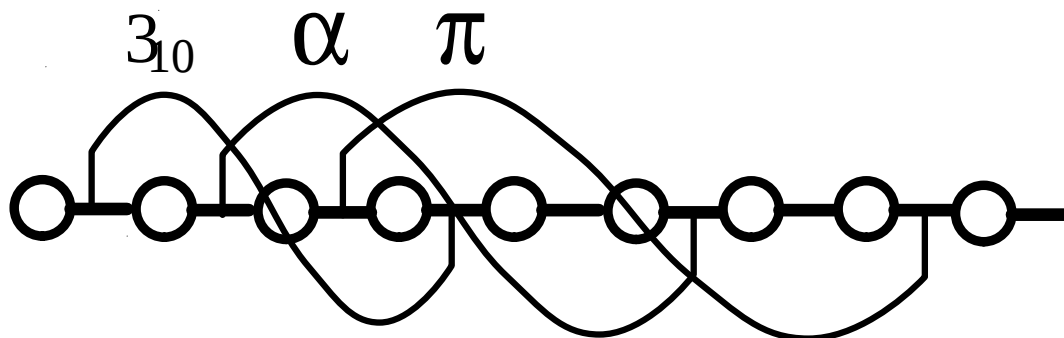


Figure 16: **A difficult secondary structure assignment.** A schematic polypeptide chain is shown with three local hydrogen-bonds (curved lines) at three different separations: $N_2 \cdots O_4 = 3_{10}$, $N_3 \cdots O_6 = \alpha$ and $N_4 \cdots O_8 = \pi$. As these segments overlap, it would be difficult to make a clear call: particularly for residue four which is incorporated in all three helix types.

of the three inertial axes are given by A, B and C (in descending order), then for a good linear structure the ratio $A/(B+C)$ will be large. This ratio can be calculated for all segment sizes at all residue positions and the problem is then just to find the optimal combination of segments.

To make the calculation more equivalent over β -strands and α -helices, the protein structure was initially smoothed by averaging successive triples of α -carbons, as described elsewhere (Taylor, 1999c). This reduces regions of α -helix and β -strand to roughly linear segments which will then have comparable ratios when calculated using the above formula. While not strictly necessary, this results in a more ‘even-handed’ treatment in the further processing of the segments described below. No smoothing or inertial ratios, however, were calculated over chain breaks.

10.1.3 Dynamic programming solution

As with many problems that incorporate a linear-ordering constraint, the optimal solution (for a given scoring scheme) can be found by the application of the dynamic programming algorithm. (See Section 3.4). The approach to the current problem follows in a way similar to the definition of trans-membrane segments (Jones *et al.*, 1994).

The basic working construct is a matrix of which the dimensions are sequence position against window size. For each value of these components, the inertial

ratio $A/(B+C)$ can be calculated. Generally, long thin structures will have a high value but so also will small structures: indeed, for the trivial case of two residues, the value will be infinite but it will also tend to be higher for smaller structures. To prevent the unwanted solution of a series of very short segments, not only was a minimum (total) segment size of five set, but the bias was tipped towards larger segments by assigning them the sum of all the values of all their sub-segments. This can be calculated quite efficiently simply by summing the raw ratio scores for a given segment with the summed values below it in the score matrix. Defining the window at residue i as $i \pm m$ (that is a window of size $2m + 1$), then:

$$s_{i,m} = r_{i,m} + r_{i-1,m} + s_{i-1,m-1} + s_{i+1,m-1} - s_{i,m-2} - am - b \quad (2)$$

where s designates the summed scores and r is the raw ratio of the inertial axes for the current window on residue i . Clearly, the score matrix can be filled recursively, as is the score matrix in sequence alignment. (See Figure 9 and Section 3.4). The subtraction of the terms am and b in Equⁿ. 2 can be chosen to prevent the summed score from monotonically increasing with window size. They are somewhat equivalent to the use of the two gap-penalties in sequence alignment (Section 3.4) with b being a fixed penalty and a controlling the increase of the penalty with segment size.

The choice of a and b in Equⁿ. 2 controls the typical segment size: if these are zero then one big segment will be obtained dropping through a series of shorter segments with increasing a and b . This can be seen in the example in Figure 18 in which a bent helix can be defined as either one or two segments.

10.1.4 ‘Continuous’ secondary structure types

The above approach parses the protein structure into lines and each line can be characterised by the residue/length (referred to below as its residue-density). This measure is effectively equivalent to a definition of secondary structure but, unlike the definition of secondary structure based on hydrogen-bonds, it is not discrete and it is thus unnecessary to make explicit definitions of secondary structure type — so allowing more freedom for ambiguous structures (loops, 3_{10} -helices or distorted β -strands) to assume different rôles. Indeed, the problem of the pathological structure described in Figure 16 is resolved, as it becomes identified as a clearly linear segment with a residue density approximating the α -helix.

11 Ideal Forms

In domain sized units, the secondary structures are typically between 10–20 Å in length and pack at roughly 10Å apart. This makes 10Å a convenient unit in which to describe their interactions in a simplified form. Further regularity is

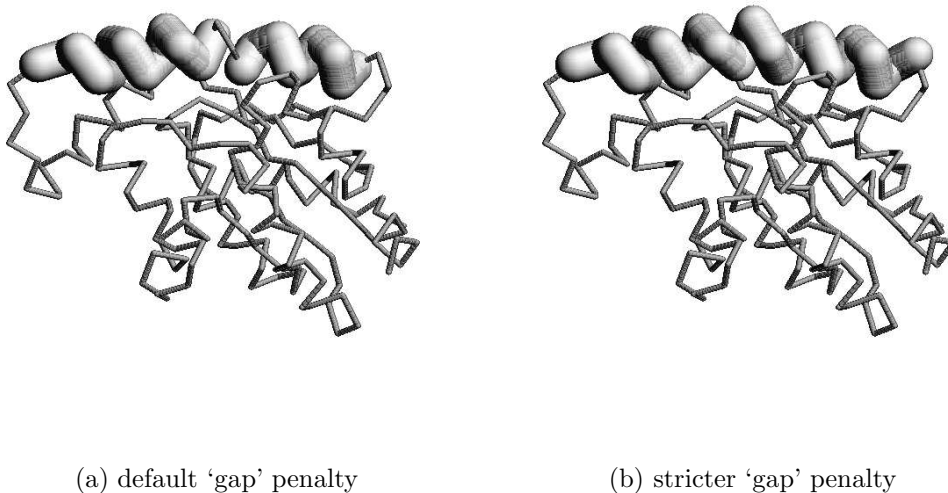


Figure 18: **Line segment variations** A small β/α protein (adenylate kinase) segmented under different 'gap' penalties. (a and b in Equ^n . 2). In the region of variation, the segment differences are emphasised using a thick line representation. (a) using the default parameters $a = b = 1$ a long helix is broken into two parts. (b) with $a = 0.5$ ($b = 1$) a single (slightly kinked) helix is selected.

introduced in the form of the β -sheet which has a strictly set β -strand spacing of (just under) 5Å. Together, these dimensions can be used to generate an idealised stick figure to represent a protein. In this section, some basic forms will be described and a method outlined which will allow them to be identified in the stick representations derived from real proteins (above).

11.1 Layer Architectures

As was described in Section 2, the units of globular proteins are secondary structures which pack together to form a hydrophobic core. Providing the protein main-chain atoms are tied-up in one of the two secondary structure types, a core can be constructed using any mix of α or β layers (Chothia and Finkelstein, 1990; Finkelstein and Ptitsyn, 1987). Seldom more than four layers are ever seen in proteins and as these can be composed of only one of two secondary structures (i.e. no mixed layers), then the possibilities are few enough to enumerate.

- 2 layers: $\beta\beta$; $\alpha\beta$; $\alpha\alpha$.
- 3 layers: $\beta\beta\beta$; $\alpha\beta\beta$, $\beta\alpha\beta$; $\alpha\alpha\beta$, $\alpha\beta\alpha$; $\alpha\alpha\alpha$.
- 4 layers: $\beta\beta\beta\beta$; $\alpha\beta\beta\beta$, $\beta\alpha\beta\beta$; $\alpha\alpha\beta\beta$, $\beta\alpha\alpha\beta$, $\alpha\beta\alpha\beta$, $\alpha\beta\beta\alpha$; $\alpha\alpha\alpha\beta$, $\alpha\alpha\beta\alpha$; $\alpha\alpha\alpha\alpha$.

(These combinations allow for reversals since proteins do not distinguish top from bottom.)

This gives 19 possible combinations, but this is an over-estimate since adjacent layers of α -helices are not always distinct. (The helices lack the strict registration imposed by the hydrogen bonding through the β -sheet.) Among these, not all possibilities are equally favoured in nature: amongst the 3-layer options, the $\alpha\beta\alpha$ combination is very widespread while in the 4-layer structures, the corresponding $\alpha\beta\beta\alpha$ structure is also encountered frequently.

11.1.1 $\alpha/\beta/\alpha$ layers

The ideal form taken to represent these structures is similar to that used previously for prediction (Cohen *et al.*, 1982) that consisted of a core β -sheet with a 20° twist between β -strands (spaced at 5Å at their mid-points). The α -helices were placed above and below this sheet using a construction that preserved the local interactions with the sheet as previously used in the construction of ideal frameworks for transmembrane helices (Taylor *et al.*, 1994b), creating a realistic staggered packing between the helices. Each helix lay, on average, 10Å above the sheet and each secondary structure was 10Å in length. (Figure 19(a)).

11.1.2 β/β layers

The model for the $\alpha/\beta/\alpha$ layer structures can also be used for stacked β proteins by neglecting the β -strands (the middle layer) and reducing the scale by half. If the outer layers (previously α -helices) are taken as β -strands then the model is a good description of two twisted β -sheet packing against each other (Taylor, 1993). This is similar to that used previously in prediction by Cohen *et al.* (1980) and more recently by Finkelstein and Reva (1991) (using a self-consistent field method). (Figure 19(b)).

Both models can be extended into a general helical structure, allowing any number of β -strands.

11.1.3 β/α -barrel proteins

A β/α -barrel structure can be constructed along the lines of a ‘squirrel’-cage (an exercise wheel more commonly used for pet hamsters) in which the β -strands are represented by the rungs around the circumference (Lesk *et al.*, 1989; Scheerlinck *et al.*, 1992). To maintain a twist between the β -strands, however, the two sides of the wheel must have a relative displacement, which is most simply made by connecting each rung not to its opposing neighbour but to a position slightly further round (Figure 19(c)).

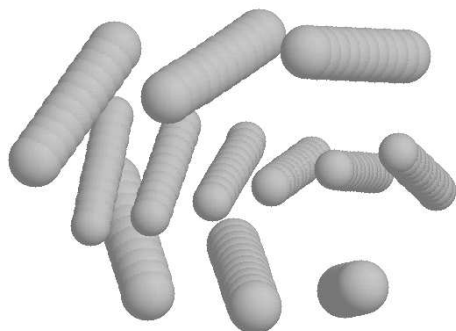
This basic model can be ‘decorated’ with α -helices in a similar way to the $\alpha/\beta/\alpha$ layers, producing a framework for the alternating β/α -barrel proteins (Figure 5).

11.1.4 All- α proteins

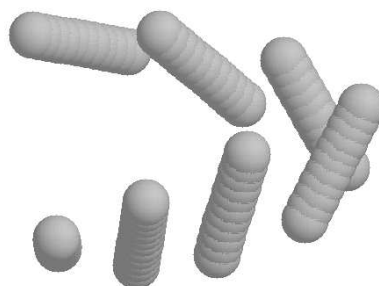
A useful model for this class was devised by Murzin and Finkelstein (1988), who, constructed idealised models for small globular proteins. If it is assumed that, to a first approximation, the core regions of α -helices are as long as they are thick, then two helices will have N- and C- terminal end-points that are equidistant both within a helix and between helices. This assumption of approximate symmetry allows very simple architectures to be constructed for bundles of packed helices in which all pairs of adjacent α -helices have equidistant end-points. This constraint, combined with the adoption of an approximately spherical shape, define a class of polyhedra that have equilateral triangles as faces and are sometimes (graphically) referred to as deltahedra. The most regular members of the class are its smallest and largest members: the tetrahedron (two helices) and at the upper end, the icosahedron (six helices). (Figure 19(d)).

11.1.5 Transmembrane models

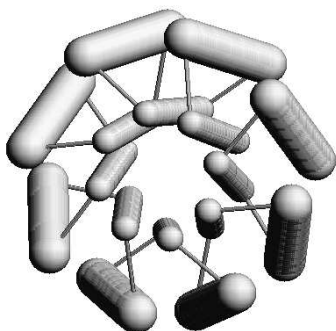
A specialised protein architecture can be found in the bundles of packed helices that typically form integral membrane proteins. Neglecting their reversed hy-



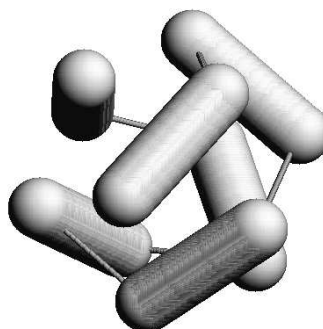
(a) $\alpha\beta\alpha$ layers



(b) $\beta\beta$ layers



(c) $\alpha\beta$ barrel



(d) all- α model

Figure 19: **Stick-figure representations.** Each of the major protein architectures are represented by their ideal ‘stick’ form. (α -helices are drawn more thickly than β -strands.) (a) $\alpha\beta\alpha$ layers. (Compare with Figure 4(b) and Figure 21). (b) Two $\beta\beta$ layers or β -‘sandwich’. Three strands pack over four — similar to the structure shown in Figure 3(a). (c) Eight-fold $\alpha\beta$ (TIM) barrel. (Compare with Figure 4(a) and Figure 5). (d) All- α model for six helices on the icosahedral frame of Murzin and Finkelstein (1988). The packing corresponds to the globin structure (Figure 3(b)). In parts *c* and *d* the fold of the equivalent proteins is shown by a fine line. The figures were produced by the program RASMOL.

drophobic polarity, these helices can also be modelled using the twisted lattice of sticks described above (Taylor *et al.*, 1994b; Bowie, 2000).

This model can also be extended into a general layer structure and can be used to overcome the limitation of the maximum number of six helices in the Murzin-Finkelstein series of deltahedra.

11.2 Stick-figure comparisons

11.2.1 Angle and Distance matching

The stick figures might be compared directly to each other using some of the structure comparison methods (described in Part II) — for example; the program **SAP** could take these data directly. However, generally, the connectivity (fold) of the ideal forms is not be specified and such a direct comparison would require testing every possible fold over the ideal form. Even for small proteins (ten segments) the number of combinations are large and quickly become excessive with larger proteins. To avoid this, the stick figures were further reduced into a matrix of pairwise line interactions. As in other similar comparison methods, such as those based on graph-matching methods (Artymiuk *et al.*, 1990) (Section 5.2) these were characterised by their distance and angle. The former was taken as the closest approach of the two line segments while the latter was the unsigned dihedral angle. These two measures are independent of line direction and so eliminate the difference between parallel and anti-parallel interactions.

Some interactions will be more important than others and this was quantified by the degree of over-lap of their line-segments. This was defined by a measure that summed a series of finely spaced lines as shown in Figure 20.

11.2.2 Finding the best match

In the **SAP** program, consecutive triples of points are taken in each structure and the similarity of the remaining points compared in the coordinate frame defined by each triple. This assessment was made on the basis of point separation and relative orientation and the best matching pairs found by dynamic-programming (Taylor and Orengo, 1989b; Taylor, 1999b). (See Part II). The current problem can be approached in a similar way, except that each triple was selected on the basis of local structural similarity with points not necessarily adjacent in the sequence. Similarly, the dynamic programming algorithm cannot be used as it assumes that the equivalent points will be in linear order. Instead the ‘stable-marriage’ algorithm (Sedgewick, 1990) was used to reconcile the matrix of conflicting preferences into a one-to-one pairwise assignment.

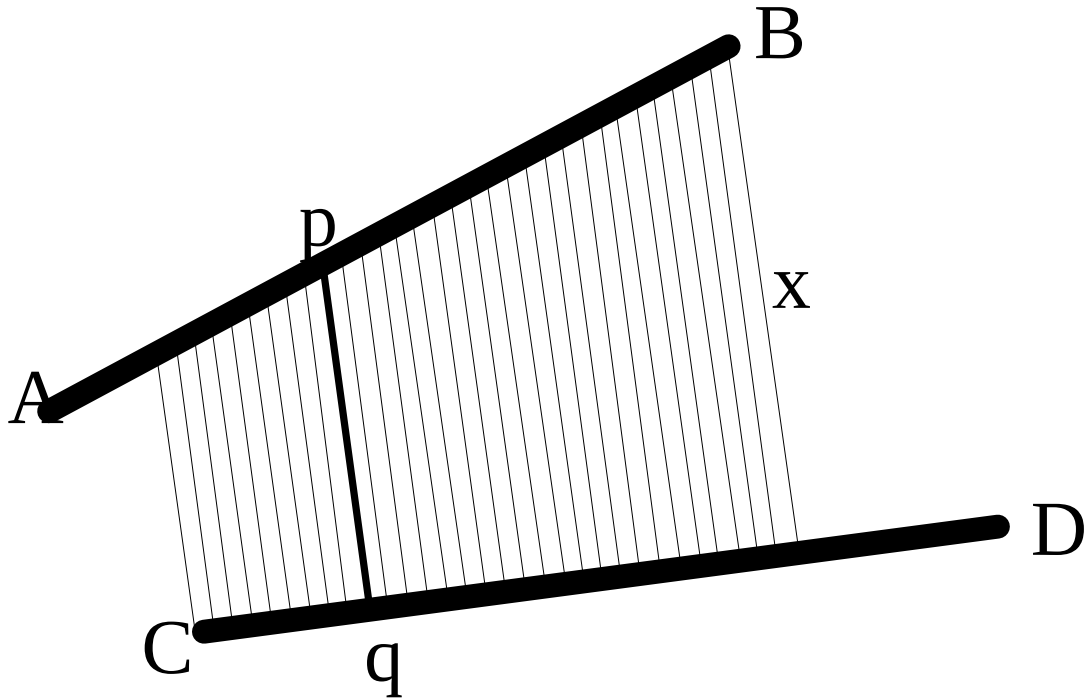


Figure 20: **Line segment overlap measure.** Two line segments corresponding to secondary structure elements are shown (**A–B** and **C–D**) as thick lines with their mutually perpendicular connecting line (**p** and **q**) shown at medium thickness. (This may lie outside one or both of the line segments). A series of fine lines cover the span in which the line segments overlap, the end-points of which are equidistant from their corresponding ends of the mutual perpendicular. A measure of interaction is calculated from this as a summation of the lengths (x) of these lines as: $\sum \exp(-x^2/a^2)$, where $a = 10$ is a good choice.

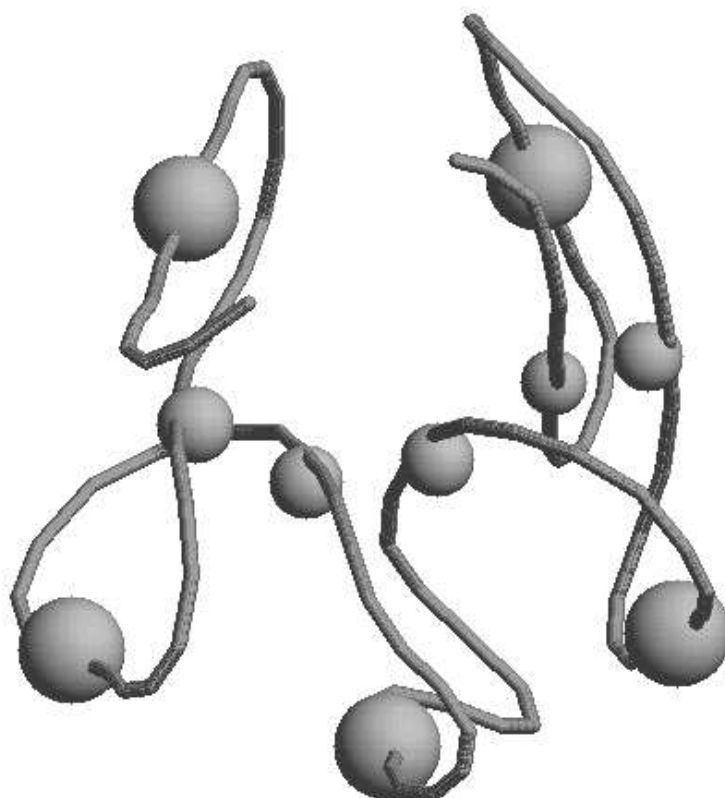


Figure 21: **Simplified representation of 3chy.** The smoothed backbone trace of the chemotaxis-Y protein is shown with the mid-points of the automatically defined line-segments shown as spheres. These have radii determined by their residue-density (see text) with the more dense segments (α -helices) appearing larger. The three-layer 2-5-3 structure can be clearly seen. (See also Figure 4(b)).

	2-5-2 (18)	3-5-2 (20)	3-6-2 (22)	3-6-3 (24)
3chy	3.305	3.260	-	-
5nul	4.002	4.471	-	-
2fcr	4.997	5.073	5.237	-
3adk	5.774	5.070	-	-
1etu	5.418	5.484	5.821	-
5p21	4.917	5.227	5.428	6.773
1kev	2.800	2.891	3.264	-

Table 1: **RMS deviations from the ideal forms** for a range of small β/α class proteins specified by their PDB codes. (See text for details). Each column gives the RMS deviation to the ideal form specified by its ‘locomotive’ class corresponding to the number of α - β - α segments in each layer. The RMS values are unweighted over all the equivalent end-points of the secondary structures, the number of which is given in parentheses at the top of each column. A dash indicates that either no solution was found by the matching program, or that which was found did not incorporate all the elements of the ideal form. Each match was examined and all were found to be a good topological match.

11.2.3 Evaluation using SAP

From the alignment of segments generated by the preceding method, it is possible to construct an ideal stick-figure with the same fold as the real protein. This reintroduces direction to the sticks and allows a direct comparison between the two structures. To make this comparison even more direct, the stick lengths of the real protein were set to the same length as their ideal counter-parts (typically 10Å). These equivalent stick figures were then passed to the **SAP** program for a full 3-D comparison. (Figure 22).

11.2.4 Nested solutions

The method described above allows a (real) protein structure to be compared to each of the ideal forms (frameworks) giving a quantified measure of each comparison. The fit of a structure to a framework will not be unique, and in general, all substructures of a framework should find a better match than the full framework itself. This is illustrated by matching a group of small $\alpha/\beta/\alpha$ type proteins against a series of nested frameworks beginning with two α -helices packed above and below a 5-stranded β -sheet. (Designated 2-5-2). The goodness-of-fit was evaluated by the RMS deviation of the real stick figure from the ideal stick figure, as calculated by the **SAP** program, based on the aligned segment end-points. (Table 1).

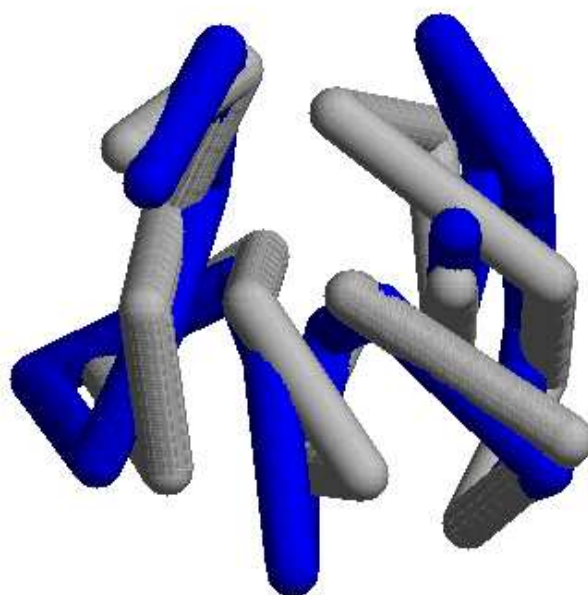


Figure 22: **Superposed stick figures of 3chy and its ideal form.** The stick figure representation of 3chy (dark grey) superposed on the corresponding stick-figure of the ideal form (light grey) is shown in the same orientation as Figure 21. The structures match with a 3.4Å RMS deviation over all 20 matched end-points.

11.3 Classification using ideal stick forms

As discussed in Section 8, with the large number of protein structures now known, it is difficult to gain an overview of their variety of forms and even more difficult to comprehend how each structure relates to its neighbours. Despite systematic attempts to instill order into this bewildering variety, the current collections (SCOP, CATH, FSSP) are all based on the pairwise comparison of protein structures. Taking this approach, the decision to group proteins together can often be arbitrary or, more cautiously, not made at all — which leads to a large number of unconnected entities.

The ability of the stick-comparison method to find solutions up to, but not beyond the core fold of the protein opens the possibility for its use as a classification tool. Given a series of ideal forms, it is necessary only to present these in order of size and select the largest solution. Unlike the visual analysis of ‘topology cartoons’, this approach is completely automatic and is focused on the well-packed core elements of the structure (which are not always obvious in topology cartoons). Finding solutions based on the core also means that two proteins can be compared even though they do not have the same overall fold. This can be done by looking back at their match to smaller ideal forms and if a common solution is found then this can be taken as a measure of relatedness.

11.3.1 A periodic table of proteins

The values in Table 1 can be presented in a more graphical form by taking the number of helices and strands in the different layers as three coordinates. The raw RMS values, will tend to be best (smallest) with the smallest structures and need to be normalised to emphasise larger structures. This can be done by turning the RMS (r) into a score (s) as:

$$s = N/(a + r) \quad (3)$$

where N is the number of matched points in the two structures (over which the RMS has been calculated) and a is a constant. When a is large, the value of r (which lies mainly in the range: 0–10) is less significant so the larger matches score most highly. With values of a less than 10, the largest match does not always have the best score and this was used in the results reported below. A graphical representation of the results for a large $\alpha/\beta/\alpha$ type protein is shown in Figure 23.

The use of such an analysis of protein structure is that it will reveal the extent to which the ideal forms are able to account for the variety of protein structure. As will be outlined in the following section, this is important for the prediction of structure from sequence. Similar approaches have also been made from the direction of a more continuous simplification of protein structure through progressive smoothing (Hinds and Levitt, 1992; Crippen and Maiorov, 1995; Özkan and Bahar, 1998).

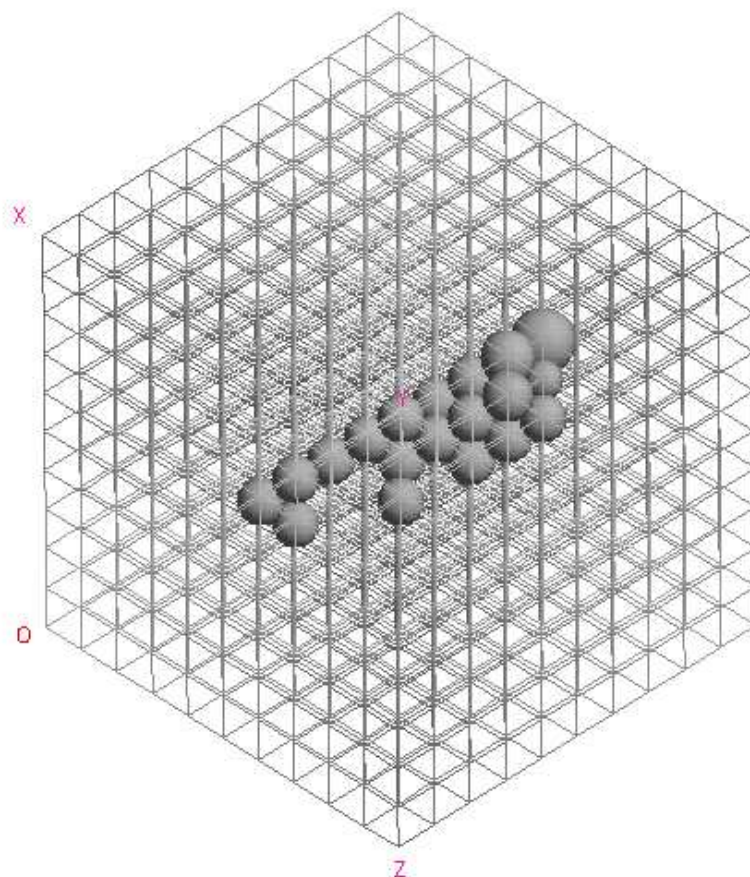


Figure 23: **Ideal substructures in a large $\beta\alpha$ protein.** Ideal $\alpha\beta\alpha$ forms were fitted to the protein (databank code **1iso**) and the number of secondary structures in the three layers plotted along each axis of a grid. The X and Z axes correspond to the two layers of α -helices, while the number of strands in the β -layer is plotted along Y (receding ‘into the page’ from the origin O). The value of s (Equⁿ. 3) is plotted for each solution as a sphere, the size of which represents the value of s . The largest (and best) solution, which has ten β -strands with four α -helices above and below the sheet, lies to the upper-right of the figure.

12 Fold Combinatorics

At the moment, the most successful prediction schemes are based on comparison of a sequence to known structures using methods such as threading (Jones *et al.*, 1992). This approach, however, is limited by the extent of the known folds and cannot, inherently, be used to predict a completely novel fold. This limitation could be overcome if the sequence were to be compared not to known folds but to idealised folds and, given a complete range of ideal forms, the problem becomes one of matching a sequence to all possible windings over each framework. This, in turn, requires generating all tracings over an idealised framework in which the path does not cross or pass through the same point twice.

Considering the finite models of Murzin and Finkelstein (1988), the register of the sequence on the framework is set by the secondary structures with each structure being placed on alternate edges of the polyhedron as the winding progresses. Computationally, this can be achieved by the application of a recursive routine which chooses a path from each node until there is no further secondary structure units or a dead-end is encountered. On each of these conditions the procedure ‘back-tracks’ to the preceding node and takes an alternative path. Exhaustive application of this procedure eventually enumerates every possible path. The introduction of distortions into these polyhedra (by displacing vertices) has also been considered (Lou *et al.*, 1993).

On the icosahedral model of six helices, there are 1264 distinct paths (the smaller but less symmetric model for five helices can generate slightly more). This can be contrasted with the alternative approach (applied to the same size of problem) of simply adding one helix onto another and allowing the fold to grow through accumulated pairwise interactions. This generates many millions of possibilities (Cohen *et al.*, 1979) most of which infringe obvious steric constraints that are never encountered when the chain is constrained to an idealised framework.

12.0.2 Motif incorporation

The possible structures generated by an unconstrained combinatoric trace over all possible windings can be greatly reduced if a distance constraint can be placed on even a pair of structural elements. In a previous study on myoglobin (Cohen and Sternberg, 1980), the constraint implied by haem binding was imposed after relatively detailed models had been built. However, it is more cost effective to apply any such constraints at an early stage. This might be done during the search over the tree of possibilities and if a forbidden pairing is encountered then all remaining combinations following that node on the tree can be neglected. This ‘tree-pruning’ strategy is most effective when the interactions being tested are sequentially local — such as the hand of $\beta\alpha\beta$ units of super secondary structure.

The single constraint of haem-binding in the globins provides a relatively

weak constraint on the possible folds, however, well defined motifs — such as the calcium binding EF-hand — can provide powerful constraints when used as a filter. For example: the protein parvalbumin contains six helices, two pairs of which constitute EF-hands. Applying these as a constraint (independently of each other) reduced the possible structures from over 1200 to three — one of which corresponded to the native fold while the other two were trivial variants (Taylor, 1991).

In the $\alpha\beta$ class, connections are also constrained by chirality and crossover. An example is shown in Figure 24 for a small protein, with databank code 3chy, used as an example frequently above (Figure 21). The secondary structures of this protein can reasonably be predicted and from considering the hydrophobicity of the β -strands, two of the five can be identified as the probable edge strands of the sheet. The topologies of all six possible arrangements of the three core strands can then easily be enumerated (Figure 24). Of these, one involves an unavoidable crossing of connecting loops while another forms a knot. Both these might reasonably be excluded (although the latter will be discussed further in Section 13.5.2).

12.1 Evaluating folds

A function is required, which, given all possible windings on an ideal framework, can recognise that which corresponds to the native fold. However, given the simplifications that are inherent in the idealised model such a function is unlikely to be reliable and attempts to specify it in terms of ‘stick’ packing have yielded little, unless specific distance or motif constraints can be incorporated (Taylor, 1991). The length of chain connecting secondary structures might be used as a constraint, but this is also not very effective, given the relatively small dimensions of the packed globule and the uncertainty in secondary structure prediction. However, the fundamental problem is that the range of interactions between pairs of secondary structures is not great since one pair of packed hydrophobic surfaces looks much like any other.

A more realistic initial step has been to apply an evaluation function to models generated from known structures. A number of methods based on empirical energy potentials allow model protein structures to be evaluated without the need to fully specify side-chain locations (e.g. Sippl (1990)). Such methods are effective at recognising protein sequences matched — or *threaded* — onto correct homologues of known tertiary structure (Jones *et al.*, 1993). In principle, it is only necessary to apply the method to matching a sequence against a sufficiently realistic representation of a combinatorially generated structures to recognise the native fold. Two practical problems barring this simple solution are the accuracy and generality of the empirical potentials used in evaluating different threadings and the realism achieved by the models generated from ‘stick’ structures.

Results based on the globins indicate that while the methods are improving,

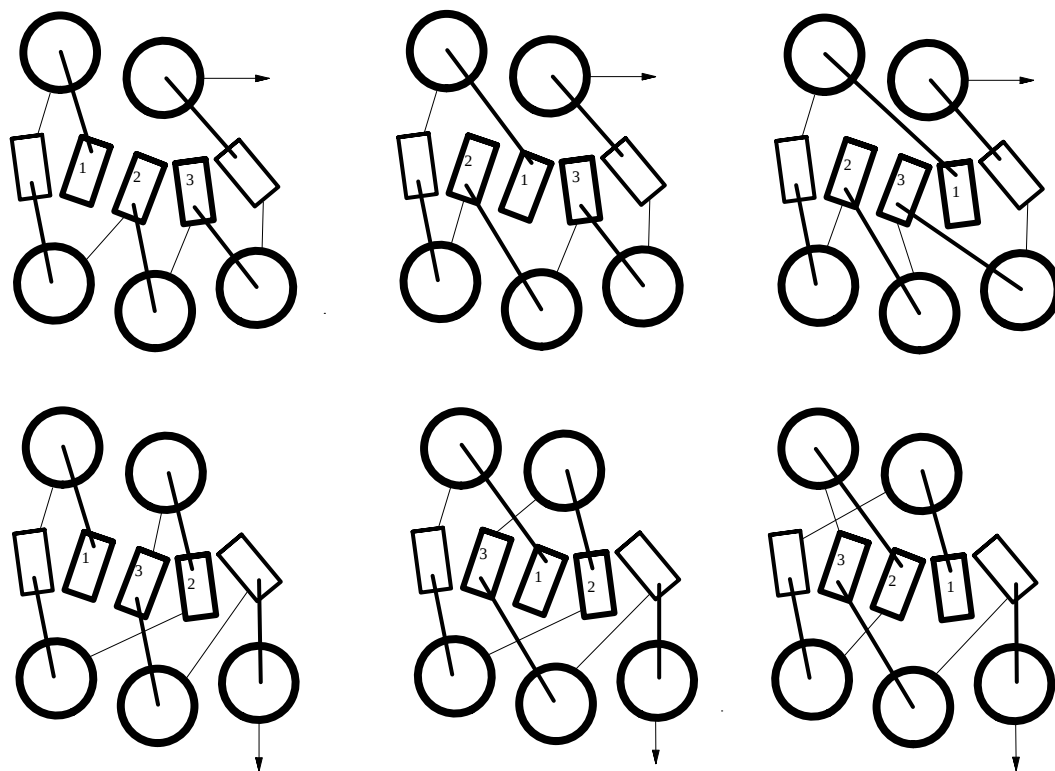


Figure 24: **Possible folds for a small $\beta\alpha$ protein.** The secondary structures for the protein 3chy can be predicted as: $\beta\alpha(\beta)\alpha\beta\alpha\beta\alpha(\beta)\alpha$, with those in parentheses being confined to the edge of the β -sheet. Labeling the remaining three core strands as 1,2,3, then their possible arrangements are: 123, 213, 231; 132, 312, 321. These are constructed in the manner of Figure 4 so as to preserve a right-handed topology of connection between strands. The 123 variant is the native ('correct') fold; folds 213, 231 and 132 infringe no known constraint (213 is found in adenylate kinase); fold 321 has an unavoidable crossover of connections and fold 312 forms a knot.

the native fold cannot yet be recognised as a unique fit. The reason for this may simply be that the ‘stick’ models are systematically different from ‘real’ proteins, so introducing an additional source of noise. Alternatively, there are many folds among the possible ‘fake’ proteins that, when viewed only at a detailed level, incorporate interactions that are more similar to the globin fold than anything encountered in the databank of ‘real’ proteins (for example; a ‘mirror-image’ globin fold). The elimination of these as candidate native folds may be impossible without full specification of (chiral) side-chain interactions.

13 Protein Topology

13.1 Introduction

Although mathematical, topology is a highly complex and abstract branch of mathematics, its roots can be traced back to simple practical problems. Knot theory, in particular, started as a subfield of applied mathematics. The first scientific application of knot theory was Gauss's work on computing the inductance of a system of linked circular wires, and Listing, who was a student of Gauss, coined the term *topology*. Since then, topological considerations have often played a rôle in theoretical problems in physics. For example; when studying the hydrodynamics of perfect fluids, Helmholtz proved that a vortex tube (a solid torus in the flow), once created, would persist in the flow forever. While his theorem illustrates the beauty and usefulness of topology in capturing the invariances in physical problems, they probably also induced Rutherford to postulate that knotted vortices in the æther might explain the different elements. Although, not supported by experiment, this intriguing theory lived long enough to give a major boost to knot theory.

As we have seen often in the preceeding Sections, the word "topology" is applied to the description of the various features in the structural hierarchy within protein molecules, from the connection patterns between secondary structure elements to the overall fold of the protein. In this Section, however, we discriminate between the 'true' topological features of proteins in the strict mathematical sense (such as intrinsic chain topology, the presence of knots and links) and the qualitative (and ill-defined) concept of the spatial arrangement of chain segments which we shall call the *fold* of the chain. This is important since in the absence of intrachain cross-links all polypeptide chains share the same intrinsic topology, namely that of the straight line segment, and are therefore indistinguishable from each other in the strict topological sense.

Before turning to proteins, we will briefly review the terminology and application of topological ideas in chemistry, giving a more general background from which applications to proteins might arise.

13.2 Chemical topology

Geometric considerations have been playing an increasingly important role in chemistry since van't Hoff postulated the tetrahedral geometry of carbon atoms in organic compounds. In fact, the development of organic chemistry provided a seemingly limitless variety of molecular shapes, the understanding of which would not be possible without the tools of topology.

Molecular structures may be regarded as graphs, where the atoms are the vertices of the graph and the edges correspond to the bonds between the atoms. Chemical graph topology has proved very useful in formalising the hitherto qual-

itative concepts of “molecular similarity” and “molecular shape”. Similarity of structures can be characterised through subgraph isomorphism matching, a technique which enables the identification of common structural motifs within molecules. (See Section 5.2 for application to protein structure comparison). The shapes of molecules can be described by various topological invariants, i.e. mappings which assign (real) numbers to graphs. Topological invariants have been used for automatic compound cataloguing and retrieval, for predicting physico-chemical properties and in quantitative structure-activity relationship (QSAR) studies.

Despite the variety of organic compounds, the overwhelming majority of them can be described by simple acyclic graphs (trees) or graphs containing a few cycles. Knots and links have not been observed and their synthesis proved difficult. The first interlocked organic molecules were synthesised as late as 1960 by Wasserman, who named them catenanes, from the Latin word *catena* (chain). These compounds contained a novel type of chemical “bond”, the *topological bond*, since they were held together by the topological arrangement of their constituent atoms, rather than by direct interatomic interactions.

13.3 Polymer topology

Natural and synthetic polymer molecules introduce an additional layer of complexity of structure which brings us closer to potential applications to protein structure. When studying the topological properties of polymers, it is often convenient to distinguish between the *intrinsic* topology and the *spatial embedding* of the structure. The intrinsic topology of the molecule is determined by the (covalent) connectivity graph of the constituent atoms, whereas the spatial embedding corresponds to the conformation of the molecule as described by the coordinates of the atoms. For example, all circular polymers have the intrinsic topology of a closed circle, but the spatial embedding of an unknotted circle is different from that of a knotted one. Conformational changes which do not require the making and/or breaking of chemical bonds are considered topologically equivalent, in line with the conventional definition of topological transformations which allow continuous deformations but no “cut-and-paste” operations.

The intrinsic topologies of polymers can be divided into a small number of major structural classes which will be discussed below. It must be noted, however, that the topology of a given molecule depends on the definition of the underlying molecular graph. In the following, we will investigate polymers at “low resolution”, by constructing molecular graphs where the nodes correspond to monomers and the arcs to bonds between monomers, thus ignoring the details of the arrangement of atoms within monomers. In some biopolymers, weaker interactions such as H-bonds often play a crucial role in structure formation; therefore, a distinction shall be made between covalent and non-covalent topologies.

13.3.1 Bond direction

In some polymers, including proteins, it is possible to assign a direction to the bonds linking the monomers. For example, in polypeptides the —NH_2 groups of the amino acid monomers form bonds with the —COOH groups and therefore each peptide bond has an amino $\rightarrow$ carboxy direction ($\text{N} \rightarrow \text{C}$ for short). Such polymers can be represented by directed graphs in which the arcs have “polarities”.

13.3.2 Linear polymers

The spatial embedding of all linear polymers are topologically equivalent since even the most tangled conformations can be transformed into a straight line by pulling the chain at one end until the whole string “flows” out smoothly. This theoretical assertion sometimes seems to contradict sharply with the practical experience concerning “knots” on ropes and tangled telephone cords, as well as folded polypeptide chains, which at first sight do not resemble straight lines at all.

The apparent inadequacy of the topological approach to describe these situations (which are directly related to the application to protein structure) can be rationalised by observing that topology concerns itself with the existence of transformations which do not change abstract properties: while the nature of the physical forces determining the conformation of a protein or a telephone cord influences the probability with which these transformations occur. However, as we shall see below, polymers with linear covalent connectivities often exhibit more complex intrinsic topologies when weaker inter-monomer interactions are taken into account, thus enabling the construction of non-trivial topological models.

13.3.3 Branching polymers

The connectivity graphs of branching polymers are trees, i.e. acyclic graphs in which there exists only one path between any two nodes. Branching polymers can also be directed if the linear branches are made up by “head-to-tail” polymerisation. At branching points, the monomers should be at least trifunctional, which is the most common case. Similarly to linear polymers, branched polymers cannot have knots or links. Natural branched polymers can be found among polysaccharides, the properties of which can be manipulated by controlling the degree of branching during synthesis.

13.3.4 Circular polymers

Circular polymers, which have the intrinsic topology of a closed loop, are particularly interesting because they can be embedded into space as knots. Also, two or more loops can be linked, giving rise to an additional topological variety. The

most important circular polymers can be found among nucleic acids. In particular, the study of topological transformations of double-stranded circular DNA molecules initiated the development of the whole field of biochemical topology (Cozzarelli and Wang, 1990).

13.4 True Topology of Proteins

13.4.1 Disulfide bridges

The sulfhydryl groups in the cysteine side chains can form disulfide bridges in an oxidative reaction. As opposed to peptide bonds, the disulfide bridges are symmetrical and therefore the covalent connectivity graph of a polypeptide with disulfide bonds can be represented by a partially directed graph. The closure of disulfide bonds creates cycles in the connectivity graph and can generate complex embedding topologies. Although the majority of such bonds form simple local connections in the sequence (Thornton, 1981) the possibility of interesting topologies has been a topic of study and speculation since the earliest days of structural work on proteins (Kauzmann, 1959; Sela and Lifson, 1959).

Crippen (1974,1975) analysed the chances of finding a knotted topology in protein chains that had been cross-linked by disulphide bridges. He simulated protein folds of different lengths as a random self-avoiding walk on a cubic lattice and then counted the knots formed. This was done in a largely automated method using an approach similar to Reidemeister moves (Adams, 1994) to reduce the complexity of the 2D projection. The chance of a knot being formed was low, at around 3% for a protein of length 128 residues but none were seen in the few multiple disulphide linked structures known at the time (Crippen, 1974). This work was further extended through simulations that incorporated the sequence (cysteine positions) of the known proteins but these more realistic simulations again suggested that proteins appeared to be "avoiding" knotted topologies. Probably, it was speculated for entropic reasons (Crippen, 1975).

On a more symbolic level, Klapper and Klapper (1980) analysed the chance of obtaining a non-planar graph in the disulphide bonded protein chain. This is a graph that cannot be drawn in 2D and is the minimal requirement for what would be considered a knotted configuration (although the Klappers used the less restrictive term of "loop-penetration"). The chance of obtaining a non-planar graph clearly increased with the number of disulphides and again the results suggested a greater chance of non-planar topologies than was later found in known protein structures. Their approach had the advantage that the disulphide bonding pattern can be known from chemical sequencing studies without having the full 3D atomic structure. However, while one case was substantiated by the 3D structure (scorpion neurotoxin) their prediction for a knot in colipase was not found in the 3D structure (implying an error in the chemical bond assignment). This was later analysed more fully by Mao (1993) along with the addition of another example

in the light-chain of the protein methylamine dehydrogenase.

The number of possible disulfide bonding arrangements in a polypeptide chain can be determined from the following formula:

$$\alpha(M, n) = {}_M C_{2n} P(n) = \frac{M!}{2^n n! (M - 2n)!}, M \leq 2n \quad (4)$$

where n is the number of disulfide bonds and M is the number of cysteines in the chain (Sela and Lifson, 1959). Within these patterns, Benham and Jafri (1993) defined the special cases of *symmetric* and *reducible* patterns. A pattern is symmetric if its mirror image (with the backbone direction reversed) has the same disulfide connections as the original, and reducible if it gives rise to two separate non-trivial subpatterns when cut once somewhere along the backbone. The same authors also carried out a statistical survey of the structure data base to assess the probabilities with which the various subpatterns occur. Symmetric and reducible patterns were observed with a much higher frequency than which was expected from theoretical studies of random disulfide bond formation (Kauzmann, 1959; Crippen, 1974). However, the limited size and the bias of the database did not allow for an analysis of statistical significance.

The non-trivial intrinsic covalent topologies generated by disulfide bonds may give rise to various interesting embeddings (knots and links). However, neither true knots nor links were found in database searches (Benham and Jafri, 1993), indicating that non-trivial disulfide bond topologies must be extremely rare if not absent among native proteins. The absence of true links in which the loops share no common backbone segment is all the more puzzling because pseudolinks, *i. e.* interpenetrations of chain segments in which the loops formed by disulfide bonds share common parts of the backbone, have indeed been observed in proteins (Klapper and Klapper, 1980; Kikuchi *et al.*, 1986; Mao, 1989; Le Nguyen *et al.*, 1990). However, pseudolinks are topologically not equivalent to true links as can be shown by suitable continuous deformations, and their linking number is zero.

From Crippen's work, the probability of a disulfide loop participating in a true link was about 0.15. This means that well over 250 true links could be expected to occur in a database containing 2,487 disjoint disulfide loops; however, none were found Benham and Jafri (1993). This absence of true links is very unlikely to have happened by chance since the proportion of reducible bond patterns¹⁰ is larger than that was expected from probabilistic considerations. Knots were also absent from the database, although Crippen's model estimated a 4% probability for knot formation in average proteins and the probability was found to increase with the chain length. These observations suggest that some feature of protein folding works against the formation of non-trivial topologies. It is sometimes argued

¹⁰These can be considered a prerequisite for link formation but to be precise, the two loops that link do have to be disjoint, since there could be other loops spanning the interval between them and this arrangement could form a true link without being reducible.

that loop penetration is hindered by stereochemical constraints in polypeptides; however, penetration is not a prerequisite of disulfide knot formation since these can be constructed by appropriately twisting hairpin loops and then linking them together. If protein folding occurs in a hierarchical fashion, with small local regions of the chain folding first and then these regions packing together, coupled with disulfide bond formation at the early stages (and consequently restricted to happen within the local folding units), then the relative abundance of reducible disulfide patterns and the scarcity of knots and true links could be explained. However, neither the current theoretical knowledge nor the available experimental information is sufficient to decide the correctness of this assumption.

13.4.2 Other cross-links

There is a very wide variety of post-translational modifications made to proteins and many of these introduce cross-links, either through direct enzymatic modification of the protein itself, or through the binding of metals and other cofactors. (See Kyte (1995) for details). Many of these modifications link two sites on the protein and so open the possibility for the creation of linked loops and knots. A wide variety of these have been analysed by Liang and Mislow (1994a/b, 1995) .

13.5 Pseudo-Topology of Proteins

Without covalent cross-linking, the formal topological analysis of proteins is greatly limited. Some further progress can be made, however, if the strict covalent bonding criterion for graph connectivity is relaxed. This can be progressed in two directions: either by considering weaker bonds, such as hydrogen-bonds as valid links, or more simply, by joining the two ends of the protein chain to form a circle.

13.5.1 Topology of weak links in proteins

In their analysis of disulphide bonded proteins (above), Klapper and Klapper (1980) introduced the idea of “loop penetration”, being a less restrictive interpretation of a knotted state defined by the covalent network being non-planar. This approach was generalised by Connolly *et al.* (1980) who defined cross-links to be any pair of α -carbon atoms that came within 7Å. (This includes all disulphide links). This looser definition encompassed a correspondingly wider variety of proteins and topological features which were referred to generally as “threaded loops”. Some folding ideas of how such features could arise were discussed.

A further generalisation of this approach is to consider all distances in proteins as potential ‘cross-links’. Each link can be characterised by the number of residues that have been ‘short-circuited’ by the connection and this value plotted against the two residue positions. The resulting plots, while similar to the Phillips (1970)

distance plots, give a good impression of the sequential packing order of the protein (Aszódi and Taylor, 1993).

13.5.2 Topology of ‘circular’ proteins

Given a piece of string, it can usually be decided by pulling the ends whether it is knotted or not. Since we hold the ends, the string plus body combination forms a closed circle and there is no danger of untying the knot as it is pulled. One way to approach the problem of defining knots in proteins is simply to join the ends (as we do when we pick up a string). This is trivial for knots where the ends of the string are remote from the knot site — but if the ends are tangled-up together with the knot then any algorithm devised to ‘pick-up’ the ends creates the risk that the external connections might either untie an existing knot or create a new one. Fortunately, for proteins, the ends of their chains (being charged) tend to lie on the surface of the structure (Thornton and Sibanda, 1983) and so can often be joined unambiguously by a wide loop. Usually, this was done by extending the termini to ‘infinity’ in a direction away from the centre of mass but the closer the termini lie to the centre of the protein, then the more arbitrary this direction will become.

With the two ends of a protein chain joined, the resulting circle can then be analysed using ‘proper’ knot theory. This approach was originally based on representing the cross-overs in a two-dimensional projection of the protein in a matrix. For example; if each section between crossings (specifically just under-crossings) is given an index, then for each crossing, we have a pair of indices and the type of crossing (effectively, left or right handed) can be entered into a matrix. The properties of such a matrix were analysed by Alexander who found that a polynomial of the matrix captured an invariant property that corresponded to its state of knotting. This was not a unique mapping as some knots could not be distinguished, but with further refinements, the distinction of knots was improved. Further progress came largely from the work of Vaughan Jones, who recast the problem as a series of ‘edit-operations’ on the knot (called skein moves), that gradually reduce the knot to a trivial form. These are ‘recorded’ in an algebraic way and also generate an answer in the form of a polynomial. The current and most powerful refinement of this approach is referred to as the HOMFLY polynomial — after the initials of the authors who developed it. (See Adams (1994) for a more complete history).

Unlike DNA, protein chains are very short (relative to their bulk) and the range of features cannot be expected to be very great. Rather than finding complex linked chains or different knot topologies (as in DNA), it is rare to find a protein chain that can even be considered as a knot. Until recently, the few folds that were reported to be knotted (without considering post-translational cross-links) have one end of the chain barely extending through a loop by a few residues and all of these form simple trefoil knots (Mansfield, 1994; Mansfield,

1997). The ‘best’ knot reported so far¹¹ requires ten residues to be removed before it becomes unknotted (Takusagawa and Kamitori, 1996).

13.5.3 ‘Topology’ of open chains

A way to avoid the unsatisfactory step of projecting the termini of the protein chain to ‘infinity’, is to reverse the operation and shrink the rest of the protein. This can be done gradually through repeated local averaging: in a chain of length N consisting of a set of coordinate vectors a ($a_1, a_2, \dots, a_N$) representing the α -carbon of each residue, each position a_i can be replaced by the average of itself and its two neighbours;

$$a_i^{t+1} = (a_{i-1}^t + a_i^t + a_{i+1}^t)/3, \quad \forall i, 1 < i < N, \quad (5)$$

where t marks the time step in the iteration. To avoid the chain passing through itself (an undesirable property for topological analysis), each move ($a_i^t \rightarrow a_i^{t+1}$) was checked to ensure that the two triangles formed by the points $\{a_{i-1}^t, a_i^t, a_i^{t+1}\}$ and $\{a_{i+1}^t, a_i^t, a_i^{t+1}\}$ were not intersected by any other line segment in the chain. If they were, then the new position (a_i^{t+1}) was not accepted.

Repeated application of this smoothing function eventually shifts all residues towards the line connecting the two termini — unless there is a ‘knot’ in the chain as this cannot be smoothed away. In theory, this simple algorithm is sufficient to detect knots in an open chain (and is equivalent to what happens in ‘real-life’ when we pull a string tight) but, just as in ‘real’ life, the resulting knots end-up very small. Indeed, in practice, the knots can become so small that the numerical accuracy of the computer is insufficient to perform the necessary topological checks and, in a numerical equivalent of quantum tunneling, the knots become undone. This was avoided by representing each line between residues by a tube $0.5\text{\AA}$ in radius.

In practice, the test for colinearity was not made at the end but an equivalent test was made to every triple of consecutive points as the smoothing progressed. When three points were close to colinear (their cosine was less than -0.99) then the middle point was removed (providing the thin triangle formed by the three points was not intersected by any other line). In addition, where the outer two came very close (specifically, fell within the tube diameter) then the middle point was also removed. This not only improved execution time but led to an even simpler test for knots as any chain that can be reduced to just its two termini is not knotted. Chains with more than two residues remaining are either knots or tangles in which a group of moves have become ‘grid-locked’ (like ‘rush-hour’ traffic at an intersection). This latter condition was eased (but not completely eliminated) by making a slight reduction in the tube diameter any time the chain

¹¹In the preparation of this work a fresh search was made for knotted proteins and a deeply knotted example was discovered which will be described below in detail.

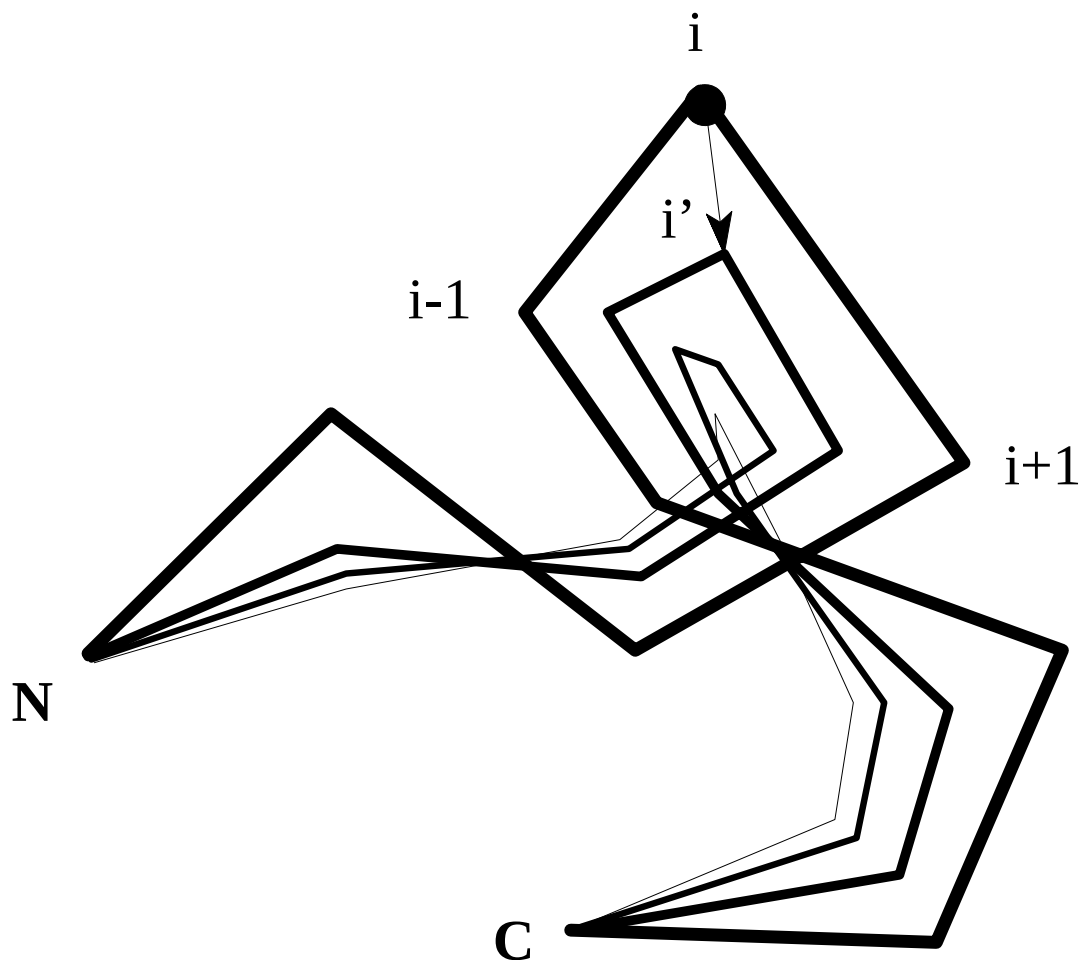


Figure 25: **The basic chain smoothing algorithm.** Protein chains are drawn schematically as lines connecting the central carbon atom in the backbone of each residue unit running from the amino (N) terminus to the carboxy (C) terminus. Beginning at the second residue, for each residue point (i) in the starting conformation, the average coordinate of i , $i-1$ and $i+1$ was taken as the new position (i') for the residue. This procedure was then repeated, and the results of this are progressively smoother chains, shown as a series of feinter lines. Note that the termini do not move. With each move, it was checked that the chains did not pass through each other. This was implemented by checking that the triangles $\{i'-1, i, i'\}$ and $\{i, i', i+1\}$ (dashed lines in the Figure) did not intersect any line segment $\{j'-1, j'\}$ ($j < i$) before the move point or any line $\{j, j+1\}$ ($j > i$) following.

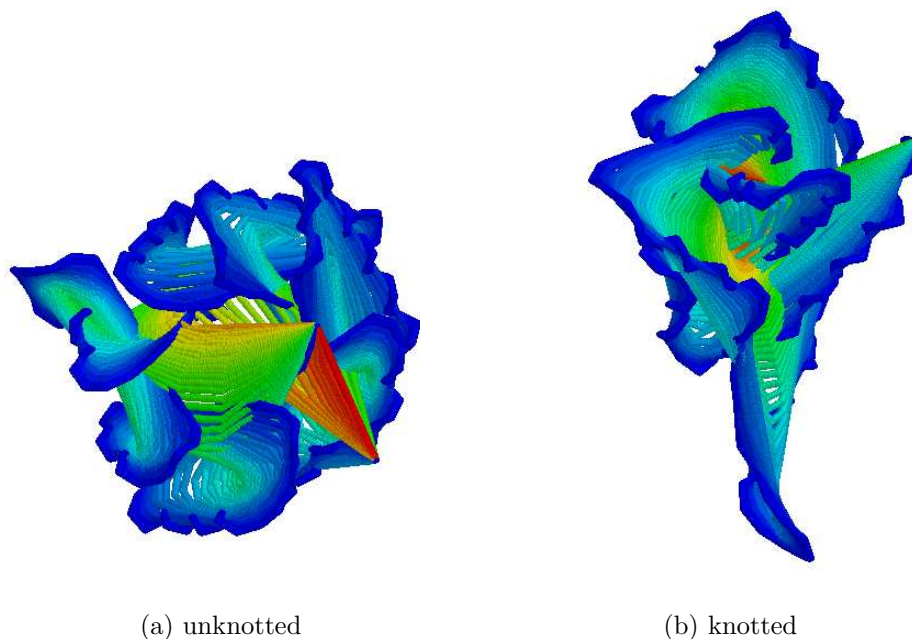


Figure 26: **Smoothed protein structures.** Applying the smoothing algorithm described in the text (also Figure 25) to protein structures produces a series of increasingly smoothed chains, coloured from blue to red in the Figures. (For clarity, the native starting structure is not shown). (a) applied to a protein that has no knots (triosephosphate isomerase, [1tph1]) results in a straight line joining the termini. To reach this stage took 52 smoothing iterations. (b) applied to the knotted protein (the carboxy-terminal domain of acetohydroxy acid isomeroreductase, [1yveI]), a straight line is never attained and a small knot remains deep in the core part of the protein. This is shown in isolation in Figure 27(a).

became stuck. Most chains of a few hundred residues are reduced to their termini in around 50 iterations. If by 500 iterations a chain was still not reduced to two points, then the resulting configuration was analysed in more detail. (Figure 26).

Importantly for proteins, the algorithm is not sensitive to the direction of projection of the termini and can therefore be used to define the exact region of the chain that gives rise to the knot. This allows knots in proteins to be characterised by how deep they lie: specifically, number of residues that must be removed from each end before they become free.

As the termini are now well separated from the knot-site, they can be unambiguously joined and analysed as a ‘proper’ circular knot. This might be done using one of the knot-invariant polynomials (discussed above). However, the few knots encountered in proteins are so simple that they do not require any sophisticated analysis and furthermore, from a theoretical perspective, not only are protein knots directional but also they have a unique break-point (between the termini)

protein	code	length	knot	core	depth
acetoxy acid isomerase	1yveI	513	U	245–444	17220
carbonic anhydrase IV	1zncA	262	A	31–261	64
ubiquitin YUH1-UBAL	1cmxA	214	H	5–210	30
VP3 core protein (bluetongue virus)	2btvB	885	L	203–879	1632
S-adenosylmethionine synthetase	1fugA	383	A	10–276	1070
carbonic anhydrase	1kopA	223	A	39–223	40
carbonic anhydrase I	1hcb	258	A	28–256	87
carbonic anhydrase V	1dmxA	237	A	5–210	22

Table 2: **Knots found in proteins.** The knots were characterised by the binary string formed by the handedness of their successive crossovers. For example, the crossovers with hands LLRR (L=left, R=right) makes 1100 which was then mapped to a letter by appending a leading 1 (giving 11100) and subtracting the value of the numerically lowest knot (1000) giving the codes: A = right-handed trefoil (RRR), H = left-handed trefoil (LLL), L = figure-of-eight (RRL), U = figure-of-eight (LLRR). The core of the knot was determined by a series of deletions from each terminus to find the smallest region that remained knotted under application of the method described in the text. To summarise this range, the product of the number of residues that must be deleted from the ends to free the knot is tabulated under 'depth'. Note that while the two trefoils (A and H) are distinct even as circular knots, the two figure-of-eight knots (U and L) can be created by introducing different break-points in a circular knot.

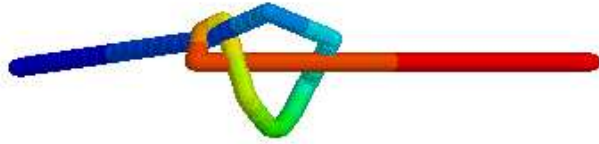
which is not taken into consideration by any of the polynomial forms. As a working tool, a simpler method was adopted to characterise these open knots based on the Dowker knot notation (Adams, 1994). In this, each crossover in a two-dimensional projection of the knot is characterised by its handedness. Beginning at the amino terminus, recording the handedness of successive crossovers as 1 or 0 generates a binary number which can be used as a reasonably unique descriptor for simple knots. To minimise the effects of projection, each knot was rotated around the axis defined by the two termini and the smallest numeric descriptor recorded.

Applying this method to a non-redundant selection of protein structures (see Table 2 for selection details) revealed a surprisingly large number of knots. A few of these proved to be unresolved tangles (including slip-knots) and some others were caused by breaks in the chain creating an unnatural short-cut. The former were all eliminated by running the program with a smaller 'tube' diameter but the latter could only be removed through visual inspection. Of the seven remaining structures (Table 2), five were right-handed trefoils including related carbonic anhydrase structures (1zncA 1kopA 1hcb 1dmxA) and the protein S-adenosylmethionine synthetase (1fugA) both of which had been identified

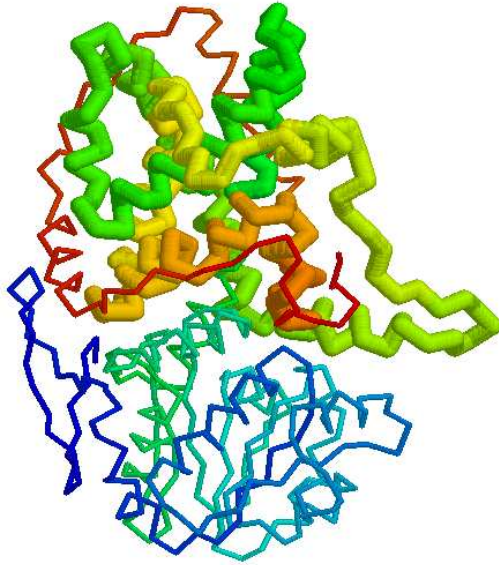
previously. In addition three novel knots were found including a left-handed trefoil in ubiquitin (**1cmxA**) and two figure-of-eight (or Flemish) knots in a viral core protein (**2btvB**) and acetohydroxy acid isoisomerase (**1yveI**). (Figure 27(a)). These latter two are of particular interest as they include an additional crossover above the trefoil and are therefore less likely to be formed by a wandering chain during folding. This was confirmed by simulation of random and semi-random compact protein-like chains in which the trefoil was by far the most common knot type. The location of the two figure-of-eight knots was determined by a series of deletions from both termini of the protein chain. This revealed that the knot in **2btvB** required the last eight residues, which is similar to the deepest trefoil knot. By contrast, the knot in **1yveI**, which is contained in the carboxy terminal domain of the protein, remained until 70 residues were deleted from the carboxy terminus and 245 residues (including a complete domain) were removed from the amino terminus. (Figure 27(b)).

It is interesting to speculate how a structure with such a deep and complex knot might fold — as it is difficult to imagine over 100 residues being ‘fed’ through a loop in a reproducible way during the folding of the protein. Clues to the folding of this protein can be found in a clear internal duplication within the domain comprising 80 residue pairs with 2.0 RMS deviation (as measured by the program **SAP** (Taylor, 1999b) over the α -carbon positions). If it is assumed that the two most deeply buried symmetrically equivalent helices initially pack together, then the remaining parts of each repeat can wrap around this core requiring only that the carboxy terminal segment can pass through the large loop between the repeats before this contracts (through the formation of α -helices) and finally packs onto the core. Following this path, the nature of the knot is determined by the chirality of the packing of the initial core helices. The symmetry in this arrangement suggests that the protein might have evolved from an exchange of structure or ‘swap’ (Bennet *et al.*, 1995) between two duplicated domains in which the first helix in the repeat has been transposed across the two-fold axis of symmetry so creating the knot. (See Section 14 for further discussion).

Integruingly, the best example of a trefoil knot (in **1fugA**) appears to have arisen in a similar manner, in which a β -strand on the edge of a sheet has been transferred from one duplicated domain to another. While it cannot be stated unambiguously that significant knots in proteins will not arise by other means, it appears that the swapping of elements of secondary structure between duplicated domains can provide a source of knotted proteins.



(a) smoothed



(b) native

Figure 27: **Knot in 1yveI**. (a) The knotted core in the smoothed representation of 1yveI (Figure 26(b)) is shown in isolation allowing the figure-of-eight knot to be seen clearly. This form was attained after 50 cycles and if continued, an irreducible core consisting of eight points was attained. (b) The backbone representation of the complete native protein structure with the minimal knotted region drawn thickened. This region is preceded by a complete nucleotide binding domain and followed by a long loop that wraps around the knotted domain.

14 Symmetry

Despite the analysis of Kendrew and colleagues when describing the first protein structure (see opening quote to Part III), it has become apparent in the intervening years, (and, hopefully, also through reading this review) that proteins are not without internal order and often symmetry. Regularities in their structure span all levels of structural organisation from the individual residue, through secondary structure (and super-secondary structure) to the overall fold of the protein. However, perhaps the greatest degree of symmetry is attained at an even higher level of the assembly of distinct protein chains (referred to as the **quaternary** structure in the hierarchy introduced in Section 2). The symmetry seen in these assemblies (which may involve one or many distinct protein types) follow general ‘rules’, exhibiting a variety of symmetry operations (Blundell and Srinivasan, 1996), but mostly simple two-, three-, and four-fold axes or extended helical arrangements — where the distinction from the large internally repeating (fibrous) protein structures becomes slight. The only symmetry operator not seen is, of course, any mirror plane.

Although fascinating, and often strongly linked to function, quaternary structure will not be pursued in this review which will stay focused on the internal organisation of proteins. Restricted to this level, a consideration of symmetry becomes a reflection of much that has been considered above — and, as such, provides an ideal topic on which to summarise and conclude.

14.1 Structural origins of fold symmetries

14.1.1 $\beta\alpha$ -class

The clear chiral preference in connections between secondary structure units — the connection β - α - β is almost never left-handed (see Section 2 and Figure 6) — can provide a strong source of symmetric structures. Imagine a protein consisting of consecutive units of α -helices and β -strands. Because local handedness is determined, all α -helices must lie on the same side of the β -sheet. However, as the α -helix is much wider than the β -strand, the structure must be curved to accommodate their differing bulk. This can result in a closed β -barrel surrounded by a ring of helices. Alternatively, if the end to which β - α units are added is reversed, the helices then fall on the opposite face of the β -sheet forming a structure with approximate two-fold symmetry. (Figure 4).

14.1.2 $\beta\beta$ -class

Equally intriguing symmetries can be found in the all- β class of structure. Typically, these are seen in structures consisting of a β -sheet (or sheets) with a closed connection forming barrel structure. If the barrel were opened-up (as in a Mercator projection of the world), the whole can be depicted in two dimensions. In

this representation, some of the chiral symmetries resemble the decorative motif commonly used in classical Greece and was accordingly named the Greek-key (Richardson, 1977). The extension of this spiral has been called a “jelly roll” (also by Richardson) and consists of eight strands in a closed barrel with two connections across top and two below.

It has been suggested that the Greek-key motif (and the jelly-roll) might have arisen from the symmetric folding of an elongated hairpin β -structure in the form of a double helix. (Figure 28)). Similar ideas can be applied to the all- α class of structure also (Finkelstein and Ptitsyn, 1987).

Some highly symmetric folds are seen in the β -trefoil and β -propeller folds. β -trefoils consist of an unusual β -sheet formed by six β -hairpins arranged with three-fold symmetry in which the connections between strands fold into three very similar units adopting a ‘Y’-like structures (Murzin *et al.*, 1992). In the larger β -propeller structure, typically six or seven β -sheets twist radially in a highly symmetric arrangement that resembles a ships propeller (Murzin, 1992)..

14.1.3 $\alpha\alpha$ -class

Folding symmetries are also found in the α/α class but their relationship to the local chiral preferences of the sub-structures are less clear. Much of the apparent symmetry within this class probably results simply from the more limited packing arrangements available with fewer secondary structures — a bundle of four or five helices will have some regularity almost no matter how they pack.

14.2 Evolutionary origins of fold symmetries

There are symmetries within some proteins that have clearly arisen through the duplication and fusion of the protein chain¹² both in the recent and remote past (McLachlan, 1972b). (For review, see Bajaj and Blundell (1984)). These include single duplication (Tang *et al.*, 1978; McLachlan, 1979; Schulz, 1980), through triple- and double-duplication (Nojima, 1987), to multiplication (McLachlan, 1983) and explosion (Higgins *et al.*, 1994). Indeed, the proteins that do not contain some indication of duplication (or pseudo-symmetry) in their structure or sequence are probably the exceptions.

More recent duplication events are often manifest as two spatially and sequentially distinct domains. However, if the original protein existed in a dimeric form, then the fused dimer can still maintain its evolved interface in the new fusion protein. This probably happened in the aspartyl proteases (Tang *et al.*, 1978). The situation can be further complicated if two symmetric parts of the

¹²Strictly, the underlying genetic code is duplicated (or translocated). Translocation requires the incorrect religation of broken double stranded DNA, while an easy route to generate duplication involves staggered (double) strand damage combined with ‘fill-in’ repair of the broken (single strand) ends before religation (Shapiro *et al.*, 1977)

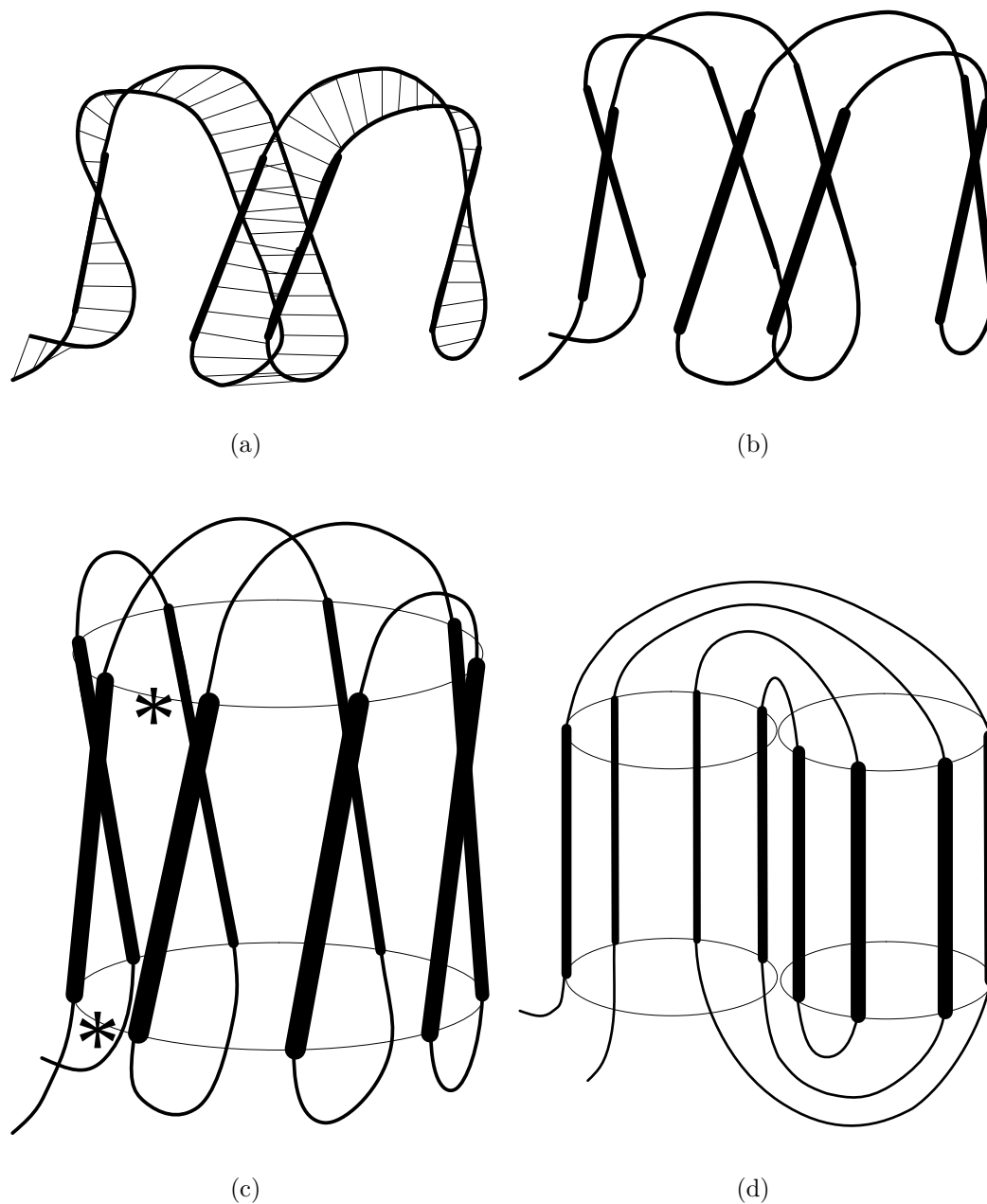


Figure 28: **Various representations of an all- β protein.** *a* Emphasising the double helical nature of the chain which may have played a rôle in folding. *b* The final double-wound structure. *c* Hydrogen-bonds between vertical strands creates a cylindrical β -sheet (arrows). *d* Opening the sheet (at the “*”s) produces a two-dimensional representation emphasising the spiral that would be seen looking down the helix axis in *a* or *b* from the left. The centre describe a *Greek Key* motif while the extended spiral is referred to as a *Jelly Roll*. Most β -sheets are less regular.

original dimer or the fusion protein have exchanged (or swapped) places (Bennet *et al.*, 1995) (see Heringa and Taylor (1997) for a review). This is probably what occurred in the knotted domain described above (Section 13.5.2). With this degree of rearrangement, the form of the original protein becomes obscured and it is often difficult to decide if the symmetry has its origins in an evolutionary event or is a consequence of purely structural pressures. Such an ambiguous example can be seen in the Rossmann fold (discussed above and Figure 4(b)).

14.3 Conclusions

We have seen that the observed shapes of proteins are a result not only of their history, but also of the physico-chemical constraints imposed by their constituent components — and it is often difficult to separate the forms imposed by these constraints from those that have been inherited. Since we, generally, have no direct access to the evolutionary history of a protein, one way to approach this problem is to quantify fully the physico-chemical constraints, then the evolutionary component (being the remainder) would similarly be known. For example, a protein function might involve a general enzymatic reaction that requires a certain juxtaposition of chemical groups (supported by a sufficiently stable framework). If it could be shown that only one protein chain fold is able to achieve this, then little evolutionary inference can be made about equivalent enzymes using this mechanism. However, if the necessary groups can be supported by, say, fifty different folds, then a group of enzymes with the same fold appears much more likely to be related.

Given the failure to predict protein structure *ab initio*, it seems unlikely that the physico-chemical constraints on structure will never be fully specified. In this situation, the most practical way forward is by inference from the sequences and structures that we can observe. Comparative analysis of these data will give indirect access to the evolutionary history of proteins and untangling these lines of descent, both within and between species, will pose a difficult challenge to the molecular evolutionist. Many of these comparisons will be difficult, if not impossible, without the aid of structural data; placing great importance on the methods of structure comparison reviewed in Part II of this work. Given sufficient sampling over this evolutionary space, we may begin to gain some idea of the structural envelope within which any given protein structure is able to be maintained. Although this characterisation maintains some affinity to Rutherford's 'stamp collecting', it is to be hoped that further structural insights will be gained along the way.

References

- Abagyan, R. A. and Maiorov, V. N. (1988). A simple qualitative representation of polypeptide chain folds: Comparison of protein tertiary structures. *J. Biomol. Struct. Dynam.*, 5(6):1267–1279.
- Abagyan, R. A. and Maiorov, V. N. (1989). An automatic search for similar spatial arrangements of α -helices and β -strands in globular proteins. *J. Biomol. Struct. Dynam.*, 6(6):1045–1060.
- Adams, C. C. (1994). *The knot book: an elementary introduction to the mathematical theory of knots*. W. H. Freeman, New York.
- Alexandrov, N. N. and Fischer, D. (1996). Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *Proteins*, 25:354–365.
- Alexandrov, N. N. and Go, N. (1994). Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Prot. Sci.*, 3:866–875.
- Alexandrov, N. N., Takahashi, K., and Gō, N. (1992). Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Molec. Biol.*, 225:5–9.
- Altschul, S. F. and Erickson, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bull. Math. Biol.*, 48:603–616.
- Artymiuk, P. J., Rice, D. W., Mitchell, E. M., and Willett, P. (1990). Structural resemblance between the families of bacterial signal-transduction proteins and of G proteins revealed by graph theoretical techniques. *Prot. Engng.*, 4(1):39–43.
- Artymiuk, P. J., Bath, P. A., Grindley, H. M., Pepperrell, C. A., Poirrette, A. R., Rice, D. W., Thorner, D. A., Wild, D. J., Willett, P., Allen, F. H., and Taylor, R. (1992a). Similarity searching in databases of three-dimensional molecules and macromolecules. *J. Chem. Inf. Comput. Sci.*, 32:617–630.
- Artymiuk, P. J., Grindley, H. M., E., P. J., Rice, D. W., and Willett, P. (1992b). Three-dimensional structural resemblance between leucine aminopeptidase and carboxypeptidase A revealed by graph-theoretical techniques. *FEBS Lett.*, 303(1):48–52.
- Aszódi, A. and Taylor, W. R. (1993). Connection topology of proteins. *Comp. App. Bio. Sci.*, 9:523–529.

- Aszódi, A. and Taylor, W. R. (1994a). Folding polypeptide α -carbon backbones by distance geometry methods. *Biopolymers*, 34:489–506.
- Aszódi, A. and Taylor, W. R. (1994b). Secondary structure formation in model polypeptide chains. *Prot. Engng.*, 7:633–644.
- Bachar, O., Fischer, D., Nussinov, R., and Wolfson, H. (1993). A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Prot. Engng.*, 6(3):279–288.
- Bajaj, M. and Blundell, T. (1984). Evolution and the tertiary structure of proteins. *Ann. Rev. Biophys. Bioeng.*, 13:453–492.
- Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., and Wilson, I. A. (1975). Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Å resolution. *Nature*, 255:609–614.
- Barton, G. J. and Sternberg, M. J. E. (1987). Evaluation and improvements in the automatic alignment of protein sequences. *Prot. Engng.*, 1:89–94.
- Barton, G. J. and Sternberg, M. J. E. (1988). LOPAL and SCAMP: techniques for the comparison and display of protein structure. *J. Molec. Graph.*, 6:190–196.
- Benham, C. J. and Jafri, M. S. (1993). Disulphide bonding patterns and protein topologies. *Prot. Sci.*, 2:41–54.
- Bennet, M. J., Schlunegger, M. P., and Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Prot. Sci.*, 4:2455–2468.
- Blundell, T. L. and Srinivasan, N. (1996). Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Nat. Acad. Sc. (USA)*, 93:14243–14248.
- Bowie, J. U. (2000). Helix-bundle membrane protein fold templates. *Prot. Sci.*, 8:2711–2719.
- Brändén, C.-I. and Tooze, J. (1991). *Introduction to Protein Structure*. Garland, New York.
- Brenner, S. E., Chothia, C., and Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, 95:6073–6078.
- Brint, A. T., M., D. H., Mitchell, E. M., and Willett, P. (1989). Rapid geometric searching in protein structures. *J. Molec. Graph.*, 7:48–53.

- Carson, M. (1991). Ribbons 2.0. *J. App. Cryst.*, 24:958–961.
- Chothia, C. and Finkelstein, A. V. (1990). The classification and origins of protein folding patterns. *Ann. Rev. Biochem.*, 59:1007–1039.
- Chothia, C. and Janin, J. (1981). Relative orientation of close-packed β -pleated sheets in proteins. *Proc. Natl. Acad. Sci. USA*, 78:4146–4150.
- Chothia, C. and Janin, J. (1982). Orthogonal packing of β -pleated sheets in proteins. *Biochemistry*, 21:3955–3965.
- Chothia, C. and Lesk, A. M. (1986). The relation between divergence of sequence and structure in proteins. *EMBO J.*, 5:823–826.
- Chothia, C. and Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harbour Symposia on Quantitative Biology*, LII:399–405.
- Chothia, C. and Murzin, A. G. (1993). New folds for all- β proteins. *Structure*, 1:217–222.
- Chothia, C., Levitt, M., and Richardson, D. (1981). Helix to helix packing in proteins. *Proc. Natl. Acad. Sci. USA*, 78:4146–4150.
- Chothia, C. (1984). Principles that determine the structure of proteins. *Ann. Rev. Biochem.*, 53:537–572.
- Chothia, C. (1992). Proteins - 1000 families for the molecular biologist. *Nature*, 357:543–544.
- Chou, K.-C., Carlacci, L., and Maggiora, G. G. (1990). Conformational and geometrical properties of idealised β -barrels in proteins. *J. Molec. Biol.*, 213:315–326.
- Clark, D. A., Shirazi, J., and Rawlings, C. J. (1991). Protein topology prediction through constraint-based search and the evaluation of topological folding rules. *Prot. Engng.*, 4:751–760.
- Cohen, F. and Sternberg, M. (1980). On the use of chemically derived distance constraints in the prediction of protein structure with myoglobin as an example. *J. Molec. Biol.*, 137:9–22.
- Cohen, F. E., Richmond, T. J., and Richards, F. M. (1979). Protein folding: Evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J. Molec. Biol.*, 132.
- Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R. (1980). Analysis and prediction of protein β -sheet structures by a combinatorial approach. *Nature*, 285:378–382.

- Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R. (1981). Analysis of the tertiary structure of protein β -sheet sandwiches. *J. Molec. Biol.*, 148:253–272.
- Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R. (1982). Analysis and prediction of the packing of α -helices against a β -sheet in the tertiary structure of globular proteins. *J. Molec. Biol.*, 156:821–862.
- Connolly, M. L., Kuntz, I. D., and Crippen, G. M. (1980). Linked and threaded loops in proteins. *19, Biopolymers*:1167–1182.
- Cozzarelli, N. R. and Wang, J. C., editors (1990). *DNA topology and its biological effects*. Cold Spring Harbor Laboratory Press, USA.
- Crippen, G. M. and Maiorov, V. N. (1995). How many protein folding motifs are there? *J. Molec. Biol.*, 252:144–151.
- Crippen, G. M. (1974). Topology of globular proteins. *J. Theor. Biol.*, 45:327–338.
- Crippen, G. M. (1975). Topology of globular proteins. II. *J. Theor. Biol.*, 51:495–500.
- Crippen, G. M. (1978). The tree structural organisation of proteins. *J. Molec. Biol.*, 126:315–332.
- Efimov, A. V. (1987). Pseudo-homology of protein standard structures formed by two consecutive β -strands. *FEBS Lett.*, 224:372–376.
- Efimov, A. V. (1991a). Structure of $\alpha - \alpha$ -hairpins with short connections. *Prot. Engng.*, 4:245–250.
- Efimov, A. V. (1991b). Structure of coiled $\beta - \beta$ -hairpins and $\beta - \beta$ -corners. *FEBS Lett.*, 284:288–292.
- Efimov, A. V. (1993). Standard structures in proteins. *Prog. Biophys. Molec. Biol.*, 60:201–239.
- Finkelstein, A. V. and Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Molec. Biol.*, 50:171–190.
- Finkelstein, A. V. and Reva, B. A. (1991). A search for the most stable folds of protein chains. *Nature*, 351:497–499.
- Fischer, D., Bachar, O., Nussinov, R., and Wolfson, H. (1992). An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dynam.*, 9(4):769–789.

- Fischer, D., Wolfson, H., and Nussinov, R. (1993). Spatial, sequence-order-independent structural comparison of alpha/beta proteins - evolutionary implications. *J. Biomol. Structure Dynamics*, 11:367.
- Fischer, D., Wolfson, H., Lin, S. L., and Nussinov, R. (1994). 3-dimensional, sequence order-independent structural comparison of a serine-protease against the crystallographic database reveals active-site similarities - potential implications to evolution and to protein-folding. *Prot. Science*, 3:769–778.
- Flores, T. P., Moss, D. S., and Thornton, J. M. (1994). An algorithm for automatically generating protein topology cartoons. *Prot. Engng*, 7:31–37.
- Flower, D. R. (1998). A topological nomenclature for protein structure. *Prot. Engng.*, 11:723–727.
- Gesteland, R. F. and Atkins, J. A., editors (1993). *The RNA world: the nature of modern RNA suggests a prebiotic RNA world*. Cold Spring Harbor Lab.
- Gibrat, J.-F., Madej, T., and Bryant, S. H. (1996). Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, 6:377–385.
- Gilbert, D., Westhead, D., Nagano, N., and Thornton, J. (1999). Motif-based searching in tops protein topology databases. *Bioinformatics*, 15:317–326.
- Gilbert, D., Westhead, D., Nagano, N., and Thornton, J. (in Press). Motif-based searching in tops protein topology databases. *Bioinformatics*.
- Godzik, A. (1996). The structural alignment between two proteins: is there a unique answer? *Prot. Sci.*, 5:1325–1338.
- Grindley, H. M., Artymiuk, P. J., Rice, D. W., and Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology*, 229:707–721.
- Hadley, C. and Jones, D. T. (1995). A systematic comparison of protein structure classifications SCOP, CATH and FSSP. *Structure*, 7:1099–1112.
- Heringa, J. and Taylor, W. R. (1997). Three-dimensional domain duplication, swapping and stealing. *Curr. Op. Struct. Biol.*, 7:416–421.
- Higgins, D. G., Labeit, S., Gautel, M., and Gibson, T. J. (1994). The evolution of titin and related giant muscle proteins. *J. Molec. Evol.*, 38:395–404.
- Hinds, D. A. and Levitt, M. (1992). A lattice model for protein-structure prediction at low resolution. *Proc. National Academy Sciences United States America*, 89:2536–2540.

- Holm, L. and Sander, C. (1993a). Globin fold in a bacterial toxin. *Nature*, 361:309.
- Holm, L. and Sander, C. (1993b). Protein-structure comparison by alignment of distance matrices. *J. Molec. Biol.*, 233:123–138.
- Holm, L. and Sander, C. (1994a). Parser for protein-folding units. *Proteins: struc. func. gene.*, 19:256–268.
- Holm, L. and Sander, C. (1994b). Searching protein structure databases has come of age. *Proteins*, 19:165–173.
- Holm, L. and Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic acids research*, 25:231–234.
- Holm, L. and Sander, C. (1998). Touring protein fold space with dali/FSSP. *Nuc. Acid. Res.*, 26:316–319.
- Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. (1992). A database of protein structure families with common folding motifs. *Prot. Sci.*, 1:1691–1698.
- Hubbard, T. J. P., Murzin, A. G., Brenner, S. E., and Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucleic Acids Research*, 25:236–239.
- Hutchinson, E. G. and Thornton, J. M. (1993). The greek key motif - extraction, classification and analysis. *Protein Engineering*, 6:233–245.
- Islam, S. A., Luo, J., and Sternberg, M. J. E. (1995). Identification and analysis of domains in proteins. *Prot. Engng.*, 8:513–525.
- Janin, J. and Chothia, C. (1985). Domains in proteins: definitions, location and structural principles. *Meth. Enz.*, 115:420–440.
- Johnson, M. S., May, A. C. W., Rodionov, M. A., and Overington, J. P. (1996). Discrimination of common protein folds: application of protein structure to sequence/structure comparisons. *Meth. Enzymology*, 266:575–598.
- Jonassen, I., Eidhammer, I., and Taylor, W. R. (1999). Discovery of local packing motifs in protein structures. *Proteins: struc. funct. gene.*, 34:206–219.
- Jones, D. T. and Taylor, W. R. (1999). Towards structural genomics for trans-membrane proteins. *Biochem. Soc. Trans.*, 26:429–438.
- Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358:86–89.

- Jones, D. T., Orengo, C. A., Taylor, W. R., and Thornton, J. M. (1993). Progress towards recognising protein folds from amino acid sequence. *Prot. Engng.*, 6 (supplement):124. (abstract).
- Jones, D. T., Taylor, W. R., and Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33:3038–3049.
- Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.
- Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, A32:922–923.
- Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, A34:827–828.
- Kajva, A. V. (1992). Left-handed topology of super-secondary structure formed by aligned α -helix and β -hairpin. *FEBS Lett.*, 302:8–10.
- Kanaoka, M., Kishimoto, F., Ueki, Y., and Umeyama, H. (1989). Alignment of protein sequences using the hydrophobic core scores. *Prot. Engng.*, 2:347–351.
- Karpen, M. E., de Haseth, P. L., and Neet, K. E. (1989). Comparing short protein substructures by a method based on backbone torsion angles. *Prot. Struct. Funct. Genet.*, 6:155–167.
- Kauzmann, W. (1959). Relative probabilities of isomers in cystine-containing randomly coiled polypeptides. In Benesch, R. e. a., editor, *Sulfur in Proteins*, pages 93–108. Academic press.
- Kikuchi, T., Némethy, G., and Scheraga, H. (1986). Spatial geometric arrangements of disulphide-crosslinked loops in proteins. *J. Compu. Chem.*, 7:67–88.
- Klapper, M. H. and Klapper, I. Z. (1980). The ‘knotting’ problem in proteins: loop penetration. *Biochim. Biophys. Acta*, 626:97–105.
- Kraulis, P. J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. App. Cryst.*, 24:946–950.
- Kuntz, I. D., Crippen, G. M., Kollman, P. A., and Kimelman, D. (1976). Calculation of protein tertiary structure. *J. Molec. Biol.*, 106:983–994.
- Kyte, J. (1995). *Structure in Protein Chemistry*. Garland Publishing, New York and London.

- Le Nguyen, D., Heitz, A., Chiche, L., Castro, B., Boigegrain, R., and Coletti-Previero, M. (1990). Molecular recognition between serine proteases and new bioactive micropoteins with a knotted structure. *Biochimie*, 72:431–435.
- Lesk, A. M. and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Molec. Biol.*, 136:225–270.
- Lesk, A., Levitt, M., and Chothia, C. (1986). Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Prot. Engng.*, 1:77–78.
- Lesk, A. M., Branden, C. I., and Chothia, C. (1989). Structural principles of α/β -barrel proteins: the packing of the interior of the sheet. *Prot. Struct. Funct. Genet.*, 5:139–148.
- Lesk, A. M. (1979). Detection of three-dimensional patterns of atoms in chemical structures. *Comm. ACM*, 22(4):219–224.
- Levine, M., Stuart, D., and Williams, J. (1984). A method for the systematic comparison of the three-dimensional structures of proteins and some results. *Acta Crystallogr.*, A40:600–610.
- Levitt, M. and Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, 261:552–558.
- Levitt, M. and Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci USA*, 95:5913–5920.
- Liang, C. and Mislow, K. (1994a). Knots in proteins. *J. Am. Chem. Soc.*, 116:11189–11190.
- Liang, C. and Mislow, K. (1994b). Topological chirality of proteins. *J. Am. Chem. Soc.*, 116:3588–3592.
- Liang, C. and Mislow, K. (1995). Topological features of protein structures: knots and links. *J. Am. Chem. Soc.*, 117:4201–4213.
- Liebman, M. N. (1982). Correlation of structure and function in biologically active small molecules and macromolecules by distance matrix partitioning. In Griffin, J. F. and Duax, W. L., editors, *Molecular Structure and Biological Activity*, pages 193–211, New York. Elsevier.
- Lou, X., Taylor, K., and Mezey, P. G. (1993). vertex mobility of polyhedra. *Bull. Math. Biol.*, 55:131–140.

- Maiorov, V. N. and Crippen, G. M. (1994). Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.*, 235:625–634.
- Mansfield, M. L. (1994). Are there knots in proteins. *Nature Struc. Biol.*, 1:213–214.
- Mansfield, M. L. (1997). Fit to be tied. *Nature Struc. Biol.*, 4:116–117. News and Views.
- Mao, B. (1989). Molecular topology of multiple-disulphide polypeptide chains. *J. Am. Chem. Soc.*, 111:6132–6136.
- Mao, B. (1993). Topological chirality of proteins. *Prot. Sci.*, 2:1057–1059.
- Matthews, B. W. and Rossmann, M. G. (1985). Comparisons of protein structures. *Meth. Enzymol.*, 115:397–420.
- Matthews, B. W., Remington, S. J., Grütter, M. G., and Anderson, W. F. (1981). Relation between hen egg white lysozyme and bacteriophage T4 lysozyme: Evolutionary implications. *J. Molec. Biol.*, 147:545–558.
- May, A. C. W. and Johnson, M. S. (1994). Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimisation. *Prot. Eng.*, 7:475–485.
- May, A. C. W. and Johnson, M. S. (1995). Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions. *Prot. Eng.*, 8:873–882.
- May, A. C. W. (1996). Pairwise iterative superposition of distantly related proteins and assessment of the significance of 3-d structural similarity. *Prot. Engng.*, 9:1093–1101.
- McLachlan, A. D. (1972a). A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr.*, A28:656–657.
- McLachlan, A. D. (1972b). Repeating sequences and gene duplication in proteins. *J. Molec. Biol.*, 64:417–437.
- McLachlan, A. D. (1979). Gene duplication in the structural evolution of chymotrypsin. *J. Molec. Biol.*, 128:49–79.
- McLachlan, A. D. (1983). Analysis of gene duplication repeats in the myosin rod. *J. Molec. Biol.*, 169:15–30.
- McLachlan, A. D. (1984). How alike are the shapes of two random chains? *Biopolymers*, 23:1325–1331.

- Mitchell, T. J., Tute, M. S., and Webb, G. A. (1989). A molecular modeling study of the interaction of noradrenaline with the beta-2-adrenergic receptor. *J. Comp. Aided Molec. Des.*, 3:211–223.
- Muirhead, H., Cox, J. M., Mazzarella, L., and Perutz, M. F. (1967). Structure and function of haemoglobin III. A three-dimensional Fourier synthesis of human deoxyhaemoglobin at 5.5 Å resolution. *J. Molec. Biol.*, 28:117–156.
- Murthy, M. R. N. (1984). A fast method of comparing protein structures. *FEBS Lett.*, 168(1):97–102.
- Murzin, A. G. and Finkelstein, A. V. (1988). General architecture of the α -helical globule. *J. Molec. Biol.*, 204:749–769.
- Murzin, A. G., Lesk, A. M., and Chothia, C. (1992). β -trefoil fold: patterns of structure and sequence in the kunitz inhibitors interleukins-1 β and 1 α and fibroblast growth factors. *J. Molec. Biol.*, 223:531–543.
- Murzin, A. G., Lesk, A. M., and Chothia, C. (1994a). Principles determining the structure of β -sheet barrels in proteins: I a theoretical analysis. *J. Molec. Biol.*, 236:1396–1381.
- Murzin, A. G., Lesk, A. M., and Chothia, C. (1994b). Principles determining the structure of β -sheet barrels in proteins: II the observed structures. *J. Molec. Biol.*, 236:1382–1400.
- Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Molec. Biol.*, 247:536–540.
- Murzin, A. G. (1992). Structural principles for the propeller assembly of β -sheets: the preference for seven-fold symmetry. *Prot. Struct. Funct. Genet.*, 14:191–201.
- Nagano, K. (1977). Logical analysis of the mechanism of protein folding: IV. super-secondary structures. *J. Molec. Biol.*, 109:235–250.
- Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Molec. Biol.*, 48:443–453.
- Nishikawa, K. and Ooi, T. (1974a). *J. Theor. Biol.*, 48:443–453.
- Nishikawa, K. and Ooi, T. (1974b). Comparison of homologous tertiary structures of proteins. *J. Theor. Biol.*, 43:351–374.

- Nojima, H. (1987). Molecular evolution of the calmodulin gene. *FEBS Lett.*, 217:187–190.
- Nussinov, R. and Wolfson, H. J. (1991). Efficient detection of 3-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. USA*, 88:10495–10499.
- Orengo, C. A. and Taylor, W. R. (1990). A rapid method for protein structure alignment. *J. Theor. Biol.*, 147:517–551.
- Orengo, C. A. and Taylor, W. R. (1993). A local alignment method for protein structure motifs. *J. Molec. Biol.*, 233:488–497.
- Orengo, C. A. and Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. In Doolittle, R. F., editor, *Computer methods for macromolecular sequence analysis*, volume 266 of *Meth. Enzymol.*, pages 617–635. Academic Press, Orlando, FA, USA.
- Orengo, C. A. and Thornton, J. M. (1993). Alpha plus beta folds revisited: some favoured motifs. *Structure*, 1:105–120.
- Orengo, C. A., Brown, N. P., and Taylor, W. R. (1992). Fast protein structure comparison for databank searching. *Prot. Struct. Funct. Genet.*, 14:139–167.
- Orengo, C. A., Flores, T. P., Jones, D. T., Taylor, W. R., and Thornton, J. M. (1993). Recurring structural motifs in proteins with different functions. *Current Biology*, 3:131–139.
- Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, 372:631–634.
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH — a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108.
- Orengo, C. A. (1994). Classification of protein folds. *Curr. Op. Struct. Biol.*, 4:429–440.
- Özkan, B. and Bahar, I. (1998). Recognition of native structure from complete enumeration of low-resolution models with constraints. *Prot. Struct. Funct. Genet.*, 32:211–222.
- Padlan, E. A. and Davies, D. R. (1975). Variability of three-dimensional structure in immunoglobulins. *Proc. Natl. Acad. Sci. USA*, 72(3):819–823.
- Pascarella, S. and Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Molec. Biol.*, 224:461–471.

- Pastore, A. and Lesk, A. M. (1990). Comparison of the structures of globins and phycocyanins - evidence for evolutionary relationship. *Prot. Struct. Funct. Genet.*, 8:133–155.
- Pearson, W. R. and Miller, W. (1992). Dynamic programming algorithms for biological sequence comparison. In Brand, L. and Johnson, M. L., editors, *Numerical Computer Methods*, volume 210 of *Methods Enzymol.*, chapter 27, pages 575–601. Academic Press Inc., N.Y.
- Phillips, D. C. (1970). The development of crystallographic enzymology. In *British Biochemistry, Past and Present*, pages 11–28, London. Biochem. Soc. Symp., Academic Press.
- Presnell, S. R. and Cohen, F. E. (1989). Topological distribution of four- α -helical bundles. *Proc. Natl. Acad. Sci. USA*, 86:6592–6596.
- Ptitsyn, O. B. and Finkelstein, A. V. (1980). Similarities of protein topologies: Evolutionary divergence, functional convergence or principles of folding? *Quart. Rev. Biophys.*, 13(3):339–386.
- Rackovsky, S. and Scheraga, H. A. (1978). Differential geometry and polymer conformation. 1. Comparison of protein conformations. *Macromolecules*, 11(6):1168–1174.
- Rackovsky, S. and Scheraga, H. A. (1980). Differential geometry and polymer conformation. 2. Development of a conformational distance function. *Macromolecules*, 13(6):1440–1453.
- Rackovsky, S. and Scheraga, H. A. (1984). Differential geometry and protein folding. *Acc. Chem. Res.*, 17:209–214.
- Rao, S. T. and Rossmann, M. G. (1973). Supersecondary structure. *J. Molec. Biol.*, 76:241–256.
- Rawlings, C. J., Taylor, W. R., Nyakairu, J., Fox, J., and Sternberg, M. J. E. (1985). Reasoning about protein topology using the logic programming language PROLOG. *J. Molec. Graph.*, 3:151–157.
- Rawlings, C. J., Taylor, W. R., Nyakairu, J., Fox, J., and Sternberg, M. J. E. (1986). Using PROLOG to represent and reason about protein structure. *Lecture Notes in Computer Science*.
- Remington, S. J. and Matthews, B. W. (1978). A general method to assess the similarity of protein structures, with applications to T4 bacteriophage lysozyme. *Proc. Natl. Acad. Sci. USA*, 75(5):2180–2184.

- Remington, S. J. and Matthews, B. W. (1980). A systematic approach to the comparison of protein structures. *J. Molec. Biol.*, 140:77–99.
- Richards, F. M. and Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: Secondary structure and first level supersecondary structure. *Prot. Struct. Funct. Genet.*, 3:71–84.
- Richardson, J. S. (1977). β -Sheet topology and the relatedness of proteins. *Nature*, 268:495–500.
- Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.*, 34:167–339.
- Richardson, J. S. (1985). Describing patterns of protein tertiary structure. *Meth. Enz.*, 115:341–380.
- Rippmann, F. and Taylor, W. R. (1991). Visualization of structural similarity in proteins. *J. Molec. Graph.*, 9:3–16.
- Rose, G. D. (1979). Hierarchic organisation of domains in globular proteins. *J. Molec. Biol.*, 234:447–470.
- Rossmann, M. G. and Argos, P. (1975). A comparison of the heme binding pocket in globins and cytochrome b_5^* . *J. Biol. Chem.*, 250:7525–7532.
- Rossmann, M. G. and Argos, P. (1976). Exploring structural homology of proteins. *J. Molec. Biol.*, 105:75–96.
- Rossmann, M. G. and Argos, P. (1977). The taxonomy of protein structure. *J. Molec. Biol.*, 109:99–129.
- Russell, R. B. and Barton, G. J. (1992). Multiple protein-sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Prot. Struct. Funct. Genet.*, 14:309–323.
- Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution. *J. Mol. Biol.*, 279:1211–1227.
- Sayle, R. and Milner-White, E. J. (1995). RasMol: Biomolecular graphics for all. *TIBS*, 20:374–375.
- Scheerlinck, J.-P. Y., Lasters, I., Claessens, M., De Maeyer, M., Pio, F., Delhaise, P., and Wodak, S. J. (1992). Recurrent $\alpha\beta$ loop structure in tim barrel motifs show a distinct pattern of conserved structural features. *Proteins*, 12:299–313.
- Schulz, G. E. (1980). Gene duplication in glutathione reductase. *J. Molec. Biol.*, 138:335–347.

- Sedgewick, R. (1990). *Algorithms in C*. Addison-Wesley.
- Sela, M. and Lifson, S. (1959). On the reformation of disulphide bridges in proteins. *Biochim. Biophys. Acta*, 36:471–478.
- Shapiro, J. A., Adhya, S. L., and Bukhari, A. I. (1977). Introduction: New pathways in the evolution of chromosome structure. In Bukhari, A. I., Shapiro, J. A., and Adhya, S. L., editors, *DNA insertion elements, plasmids and episomes*, pages 3–11. Cold Spring Harbor Lab.
- Siddiqui, A. S. and Barton, G. J. (1995). continuous and discontinuous domains – an algorithm for the automatic generation of reliable protein domain definitions. *Prot. Sci.*, 4:872–884.
- Sippl, M. J. (1982). On the problem of comparing protein structures: Development and applications of a new method for the assessment of structural similarities of polypeptide conformations. *J. Molec. Biol.*, 156:359–388.
- Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J. Molec. Biol.*, 213:859–883.
- Sklenar, H., Etchebest, C., and Lavery, R. (1989). describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Prot. Struct. Funct. Genet.*, 6:46–60.
- Smith, R. F. and Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Prot. Eng.*, 5:35–42.
- Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Molec. Biol.*, 147:195–197.
- Sowdhamini, R. and Blundell, T. (1995). An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Prot. Sci.*, 4:506–520.
- Sternberg, M. J. E. and Thornton, J. M. (1977a). On the conformation of proteins: An analysis of β -pleated sheets. *J. Molec. Biol.*, 110:285–296.
- Sternberg, M. J. E. and Thornton, J. M. (1977b). On the conformation of proteins: The handedness of the connection between parallel β -strands. *J. Molec. Biol.*, 110:269–283.
- Sternberg, M. J. E., Taylor, W. R., Nyakairu, J., Fox, J., and Rawlings, C. J. (1985). Reasoning about protein topology using the logic programming language PROLOG. *J. Molec. Graph.*, 3:108–109. (abstract).

- Subbarao, N. and Haneef, I. (1991). Defining topological equivalences in macromolecules. *Prot. Eng.*, 4:887–884.
- Subbiah, S., Laurents, D. V., and Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.*, 3(3):141–148.
- Swindells, M. B. (1995). A procedure for detecting structural domains in proteins. *Prot. Sc.*, 4:103–112.
- Takusagawa, F. and Kamitori, S. (1996). A real knot in protein. *J. Am. Chem. Soc.*, 118:8945–8946.
- Tang, J., James, M. N. G., Hsu, I. N., Jenkins, J. A., and Blundell, T. L. (1978). Structural evidence for gene duplication in the evolution of the acid proteases. *Nature*, 271:619–621.
- Taylor, W. R. and Orengo, C. A. (1989a). A holistic approach to protein structure comparison. *Prot. Eng.*, 2:505–519.
- Taylor, W. R. and Orengo, C. A. (1989b). Protein structure alignment. *J. Molec. Biol.*, 208:1–22.
- Taylor, W. R., Thornton, J. M., and Turnell, W. G. (1983). A ellipsoidal approximation of protein shape. *J. Molec. Graphics*, 1:30–38.
- Taylor, W. R., Flores, T. P., and Orengo, C. A. (1994a). Multiple protein structure alignment. *Prot. Sci.*, 3:1858–1870.
- Taylor, W. R., Jones, D. T., and Green, N. M. (1994b). A method for α -helical integral membrane protein fold prediction. *Prot. Struct. Funct. Genet.*, 18:281–294.
- Taylor, W. R. (1986a). The classification of amino acid conservation. *J. Theor. Biol.*, 119:205–218.
- Taylor, W. R. (1986b). Identification of protein sequence homology by consensus template alignment. *J. Molec. Biol.*, 188:233–258.
- Taylor, W. R. (1988). A flexible method to align large numbers of biological sequences. *J. Molec. Evol.*, 28:161–169.
- Taylor, W. R. (1991). Sequence analysis: spinning in hyperspace. *Nature*, 353:388–389. (News and Views).
- Taylor, W. R. (1993). Protein structure prediction from sequence. *Computers and Chem.*, 17:117–122.

- Taylor, W. R. (1997a). Multiple sequence threading: an analysis of alignment quality and stability. *J. Molec. Biol.*, 269:902–943.
- Taylor, W. R. (1997b). Random models for double dynamic score normalisation. *J. Molec. Evol.*, 44:S174–S180. Special issue in memory of Kimura.
- Taylor, W. R. (1998). Dynamic databank searching with templates and multiple alignment. *J. Molec. Biol.*, 280:375–406.
- Taylor, W. R. (1999a). The properties of amino acids in sequences. In Bishop, M. J., editor, *Nucleic acid and protein databases: a practical approach (second edition)*, pages 81–103. Academic Press. Chapter 5.
- Taylor, W. R. (1999b). Protein structure alignment using iterated double dynamic programming. *Prot. Sci.*, 8:654–665.
- Taylor, W. R. (1999c). Protein structure domain identification. *Prot. Engng.*, 12:203–216.
- Thomas, D. T. (1994). The graduation of secondary structure elements in proteins. *J. Mol. Graphics*, 12:146–152.
- Thornton, J. and Sibanda, B. (1983). Amino and carboxy-terminal regions in globular proteins. *J. Molec. Biol.*, 167:443–460.
- Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M. G. (1999). Protein folds, functions and evolution. *J. Molec. Biol.*, 293:333–342.
- Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Molec. Biol.*, 151:261–287.
- Vingron, M. and Waterman, M. S. (1994). Sequence alignment and penalty choice: review of concepts, case-studies and implications. *J. Molec. Biol.*, 235:1–12.
- Vriend, G. and Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins*, 11:52–58.
- Šali, A. and Blundell, T. L. (1990). Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationship through simulated annealing and dynamic programming. *J. Molec. Biol.*, 212:403–428.
- Weber, P. C. and Salemme, F. R. (1980). Structural and functional diversity in 4- α -helical proteins. *Nature*, 287:82–84.

Zhu, Z.-Y., Šali, A., and Blundell, T. L. (1992). A variable gap penalty-function and feature weights for protein 3-D structure comparisons. *Prot. Engng.*, 5:43–51.

Zuker, M. and Somorjai, R. L. (1989). The alignment of protein structures in three dimensions. *Bull. Math. Biol.*, 51(1):55–78.

ACKNOWLEDGEMENTS: David Jones and Jaap Heringa are thanked for help with some of the figures.